

Leveraging Subtle Verbalization and Speech Patterns to Help Evaluators Identify Usability Problem Encounters in Concurrent Think-Aloud Sessions

by

Mingming Fan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Computer Science
University of Toronto

© Copyright by Mingming Fan 2019

Leveraging Subtle Verbalization and Speech Patterns to Help Evaluators Identify Usability Problem Encounters in Concurrent Think-Aloud Sessions

Mingming Fan

Doctor of Philosophy

Department of Computer Science
University of Toronto

2019

Abstract

Think-aloud protocols are a highly valued usability testing method for identifying usability problems. Despite the value of conducting think-aloud usability test sessions, analyzing think-aloud sessions is often arduous. Consequently, previous research has urged the community to develop methods to support fast-paced analysis. Inspired by the research that shows subtle patterns in how we interact with other people reveal our attitudes toward them, I study subtle patterns in users' verbalizations and speech when they encounter problems in think-aloud sessions and further leverage these patterns to support the analysis of think-aloud sessions.

In this dissertation, I first survey user experience (UX) practitioners around the world to understand their practices and challenges around using think-aloud protocols. I then design and conduct three studies, each addressing the limitations of the previous one, to identify and validate the subtle patterns in users' verbalization and speech features that tend to occur when they encounter usability problems. Informed by the findings from the studies, I take the first step to designing computational methods that leverage these subtle patterns and the power of machine learning (ML) to detect the usability problem encounters. To help UX practitioners leverage ML-inferred usability problem encounters, I design and evaluate an intelligent visual analytics tool that present

UX practitioners with a timeline visualization of ML-inferred problem encounters and ML's input features among other functions. Experimental results demonstrate that ML-inferred problem encounters help UX practitioners consider problems that they might have overlooked and therefore identify more usability problems. Moreover, I offer insights into how UX practitioners leverage and perceive ML-inferred problem encounters and ML's input features in their analysis and their session review strategies (i.e., how they play, pause, rewind the recorded sessions). Finally, I highlight the promising directions to further forge a better symbiosis relationship between UX practitioners and machine intelligence when analyzing recorded think-aloud sessions.

Acknowledgments

I could not complete my Ph.D. without wholehearted and continued love and support from my family. You are the reason why I have been able to pursue my dreams determinedly and fearlessly. I dedicate this dissertation to you, my dear family.

I could not thank enough my advisor Khai N. Truong. Khai adopted me as his Ph.D. student when I was a confused student in another Ph.D. program. In addition to pragmatic skills for conducting research, Khai has also demonstrated to me how to think critically about my research. I started my Ph.D. working on sensing technologies and applications. Despite having publications, I had difficulty building upon my own research and extending it into a dissertation. Realizing my weaknesses at the time, Khai devoted a substantial amount of time and effort to meet with me as much as I needed to help me get through my setback. I still vividly remember many conversations that we had in many locations beyond the DGP lab over that period. Khai, I greatly appreciate your patience, wisdom, and guidance to help me *learn from my own failure*. It is equally, if not more, valuable to me as any concrete research skills that I've learned from you. Thank you, Khai.

I would like to thank Mark Chignell and Daniel Wigdor for being on my Ph.D. committee and offering valuable feedback that has helped me polish this dissertation. I still remember that Mark attended a local technical writers' meetup and helped me connect with others and identify potential research directions. Thank you, Mark. Daniel's feedback had inspired me to conduct the survey study and helped me better ground the dissertation. Thank you, Daniel. I would also like to thank Fanny Chevalier and Mary Czerwinski for reading my dissertation and being on my final oral examination committee. Your comments and feedback are inspirational for my future research. Thank you, Fanny and Mary.

I would also like to thank Jian Zhao for his support in the design and evaluation of VisTA and all the students who had worked with me for their contributions to part of the dissertation. Thank you to Christina Chung, Yue Li, Candice Lin, Serina Shi, Winter Wei, and Ke Wu. Also, I would like to thank Olivier St-Cyr, Zi'ang Chen, and the chairs of many UXPA local chapters for their help with advertising some of the user studies that I conducted in this dissertation. Additionally, I would also like to sincerely thank all the study participants.

Being part of the DGP lab has been an enjoyable experience. DGPers, I thank every one of you for the numerous chats, laughter, encouragement, and refreshing social events (e.g., biking), which have been imperative to maintain physical and mental health. Because of you, my Ph.D. journey has been more precious and memorable. Furthermore, I would like to thank Ravin Balakrishnan for his academic career advice, John Hancock for his professional and patient support with lab facilities, and Seongkook Heo for mutual encouragement during our academic job search.

During my Ph.D. I have also been fortunate to work with many talented colleagues and students on research projects beyond the scope of this dissertation, which have substantially extended my research scope and helped me develop collaboration and mentoring skills. Thank you to Alexander T. Adams, Yizheng Ding, Teng Han, Ying Han, Franklin Li, Jiannan Li, Weiwei Li, Zhen Li, Eric Lu, Zhicong Lu, Zhongtian Qiu, Fang Shen, Yiqing Yang, Yuhui You, Zhi Yu, and Yue Zhao.

Copyright Notice and Disclaimer

Sections of this document have appeared in publications or are forthcoming (at the time of writing). In all cases, permission has been granted by the publisher for these works to appear here. Below, the publisher's copyright notice and/or disclaimer is given, with thesis chapter (s) and corresponding publication (s) noted.

User Experience Professionals Association (UXPA)

Copyright © 2019–2020, User Experience Professionals Association and the authors. 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. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. URL: <http://www.upassoc.org>.

portions of chapter 3

Mingming Fan, Serina Shi, Khai N. Truong. Practices and Challenges of Using Think-aloud Protocols in Industry: An International Survey. *Journal of Usability Studies* (In press)

Association for Computing Machinery (ACM)

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.

Copyright © ACM 2019 1073-0516/2019 \$15.00

portions of chapters 4, 5 and 6

Mingming Fan, Jinglan Lin, Christina Chung, Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Transactions on Computer-Human Interaction*. 26, 5, Article 28 (July 2019), 35 pages DOI: <https://doi.org/10.1145/3325281>

portions of chapter 7

Mingming Fan, Yue Li, Khai N. Truong. Automatic Detection of the Encounters of Usability Problems in Think-Aloud Sessions. *ACM Transactions on Interactive Intelligent Systems* (Accepted with Major Revisions).

The Institute of Electrical and Electronics Engineers (IEEE)

Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

portions of chapter 8

Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, Khai N. Truong. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics*. (In press, to be presented at IEEE VIS 2019 *InfoVis*).

Table of Contents

Acknowledgments.....	iv
Copyright Notice and Disclaimer	vi
Table of Contents	viii
List of Tables	xiv
List of Figures	xvi
Chapter 1 Introduction	1
1.1 Research Objective	3
1.2 Contributions.....	5
1.3 Dissertation Outline	6
Chapter 2 Background and Related Work	9
2.1 Think-Aloud Protocols.....	9
2.1.1 History.....	9
2.1.2 Concurrent vs. Retrospective Think-Aloud Protocols	9
2.1.3 Comparison to Other Usability Evaluation Methods.....	10
2.1.4 Validity Confirmation with Other Streams of Data.....	11
2.2 Conducting Think-Aloud Sessions	11
2.2.1 Guidelines for Conducting Think-Aloud Sessions	11
2.2.2 Relaxed Think-Aloud Protocols	12
2.3 Classic vs. Relaxed Think-Aloud Protocols	12
2.3.1 The Effect of Interaction Between the Evaluator and the Participant	12
2.3.2 The Effect of Instructions for Requesting Verbalizations	13
2.3.3 Summary	14
2.4 Verbalizations in Think-Aloud Sessions	15
2.4.1 Three-Levels of Verbalizations.....	15
2.5 Verbalization Categorization	17
2.6 Speech Features	19
2.7 Think-Aloud Protocols Use in the Industry	20
2.8 Summary	20
Chapter 3 Practices and Challenges of Using Think-Aloud Protocols in the Industry.....	23
3.1 Introduction.....	23
3.2 Goal.....	24
3.3 Participants.....	24

3.4	Survey Design.....	24
3.5	Data Analysis.....	25
3.6	Results.....	25
3.6.1	Respondents' Profile.....	25
3.6.2	General use of think-aloud protocols	28
3.6.3	Conducting think-aloud sessions	30
3.6.4	Analyzing think-aloud sessions	34
3.6.5	Challenges with using think-aloud protocols	38
3.7	Discussion.....	41
3.7.1	General use of think-aloud protocols	41
3.7.2	Conducting think-aloud sessions	42
3.7.3	Analyzing think-aloud sessions	44
3.8	Summary.....	45
Chapter 4 Study 1: Verbalization and Speech Features and Usability Problems		47
4.1	Concurrent Think-Aloud Data Collection	47
4.1.1	Participants	47
4.1.2	Procedure	48
4.2	Analysis of Think-Aloud Sessions	49
4.2.1	Participants (Evaluators).....	49
4.2.2	Study Design	49
4.2.3	Verbalization Categorization and Speech Features Extraction.....	50
4.2.4	Tool for analyzing Think-Aloud Sessions	51
4.2.5	Procedure	54
4.3	Analysis and Results.....	55
4.3.1	Problems Identified by Usability Evaluators	56
4.3.2	Verbalization Categories and Identified Problems	56
4.3.3	Speech Features and Identified Problems.....	57
4.3.4	Qualitative Feedback	58
4.4	Summary.....	62
Chapter 5 Study 2 (Confirmation Study).....		63
5.1	Concurrent Think-Aloud Data Collection	63
5.1.1	Participants.....	63
5.1.2	Procedure	64

5.2	Analysis of Think-Aloud Sessions	64
5.2.1	Participants (Evaluators).....	64
5.2.2	Study Design	65
5.2.3	Verbalization Categorization and Speech Features Extraction.....	65
5.2.4	Tool for Analyzing Think-Aloud Sessions	67
5.2.5	Procedure	67
5.3	Analysis and Results.....	68
5.3.1	Number of Labels per Verbalization Category.....	68
5.3.2	Problems Identified by Usability Evaluators	69
5.3.3	Verbalization Categories and Identified Problems	69
5.3.4	Speech Features and Identified Usability Problems	72
5.3.5	Qualitative Feedback on the Use of Verbalization & Speech Features	73
5.4	Summary	77
Chapter 6 Study 3 (Generalization Study)		81
6.1	Concurrent Think-Aloud Data Collection	82
6.1.1	Participants	82
6.1.2	Procedure	82
6.2	Analysis of Think-Aloud Sessions	84
6.2.1	Participants (Evaluators).....	84
6.2.2	Study Design	84
6.2.3	Verbalization Categorization and Speech Features Extraction	85
6.2.4	Tool for Analyzing Think-aloud sessions in Different Study Conditions	85
6.2.5	Procedure	88
6.3	Analysis and Results.....	88
6.3.1	Number of Labels per Verbalization Category.....	88
6.3.2	Problems Identified by Usability Evaluators	89
6.3.3	Verbalization Categories and the Identified Problems	90
6.3.4	Physical Devices vs. Digital Systems	91
6.3.5	Audio vs. Video	93
6.3.6	With vs. Without Visualization	94
6.3.7	What Users Talked About When They Encountered Problems?.....	95
6.3.8	Speech Features and Identified Usability Problems (i.e., How Did Users Verbalize When Experiencing Problems?).....	97
6.4	Discussion	98




6.4.1	Physical Devices vs. Digital Systems	98
6.4.2	Audio vs. Video Modality Available to Evaluators.....	99
6.4.3	Visualization of Verbalization and Speech Features	101
6.4.4	Verbalization Category Proportions	101
6.4.5	Evaluator Effect	103
6.5	Summary	104
Chapter 7 Automatic Detection of Usability Problem Encounters		107
7.1	Research Questions.....	107
7.2	Automatic Detection of Usability Problem Encounters.....	108
7.2.1	Think-Aloud Dataset.....	109
7.2.2	Ground truth Labeling.....	109
7.2.3	Basic Transcript-based Feature Extraction	109
7.2.4	Verbalization and Speech Feature Extraction.....	110
7.2.5	ML Models.....	111
7.3	Evaluation and Results.....	112
7.3.1	The Effect of the <i>Verbalization and Speech Features</i> and the <i>ML models</i> on the Detection of Usability Problem Encounters.....	112
7.3.2	The Effect of <i>Products</i> on the Detection of Usability Problem Encounters	116
7.3.3	The Effect of <i>Users</i> on the Detection of Usability Problem Encounters	117
7.3.4	Summary of the Key Findings	119
7.4	Discussion	119
7.4.1	The <i>Verbalization and Speech Features</i> are Complementary in Detection of Usability Problem Encounters	119
7.4.2	The <i>ML Models</i> for Detecting Usability Problem Encounters Have Different Precision and Recall Tradeoff	120
7.4.3	The Types of <i>Products</i> and <i>Tasks</i> Affect the Detection of Usability Problem Encounters	121
7.4.4	The <i>User-Dependent</i> Models for Detecting Usability Problem Encounters Vary in Performance	122
7.4.5	Automatic Verbalization Category Labelling	122
7.5	Summary	124
Chapter 8 Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions		127
8.1	Research Questions.....	127
8.2	Think-Aloud Dataset and Problem Encounter Detection	128

8.2.1	Dataset.....	128
8.2.2	Data Labelling and Feature Extraction	129
8.2.3	Model Training and Evaluation	130
8.3	VisTA: <i>Visual Analytics Tool for Think-Aloud</i>	131
8.3.1	Design Principles	131
8.3.2	Session Reviewing and Problem Logging	134
8.3.3	Visualization of Problem Encounters and Features	135
8.3.4	Implementation Details	137
8.4	User Study.....	138
8.4.1	Study Design.....	138
8.4.2	Participants.....	139
8.4.3	Procedure	139
8.5	Analysis and Results	140
8.5.1	Data Capture and Analysis.....	140
8.5.2	Overview of the Quantitative and Qualitative Results	141
8.5.3	Quantitative Results	142
8.5.3.1	Problem Identification	142
8.5.3.2	Session Review Strategies	143
8.5.3.3	Questionnaire.....	147
8.5.4	Qualitative Results	148
8.5.4.1	How did evaluators use the problem timeline (i.e., ML-inferred problem encounters)?	148
8.5.4.2	How did evaluators use the feature timeline (i.e., ML’s input features)?	149
8.5.4.3	What were evaluators’ attitudes toward ML-inferred usability problem encounters?.....	150
8.5.4.4	How did evaluators deal with agreement and disagreement with ML- inferred usability problem encounters?	151
8.5.4.5	What did evaluators perceive as the limitations of ML-inferred usability problem encounters?	152
8.6	Discussion	153
8.6.1	The Effect of <i>ML-inferred Usability Problem Encounters</i> on the Evaluators’ Analysis.....	153
8.6.2	Attitudes toward ML-Inferred Usability Problem Encounters	155
8.6.3	Reliance on ML-Inferred Usability Problem Encounters	157

8.6.4	Trust in ML-inferred Usability Problem Encounters.....	158
8.7	Summary.....	159
Chapter 9	Conclusion and Future Work.....	163
9.1	Summary of the Key Takeaways.....	163
9.1.1	Subtle Verbalization and Speech Patterns Tend to Occur When Users Encounter Usability Problems in Concurrent Think-Aloud Sessions.....	163
9.1.2	Subtle Verbalization and Speech Patterns Can Be Used to Build Effective ML Models to Detect Usability Problem Encounters.....	165
9.1.3	ML-Inferred Usability Problem Encounters Can Assist UX Practitioners with Identifying Usability Problems More Effectively.....	165
9.2	Limitations and Future Research Directions.....	166
9.2.1	Further Validating the Findings with <i>More Products, More Participants, and Evaluators with Different Levels of Familiarity with Test Products</i>	167
9.2.2	Understanding the Impact of <i>Language</i> and <i>Age</i> on the Subtle Verbalization and Speech Patterns that Indicate Usability Problems.....	168
9.2.3	Understanding the Impact of Alternative <i>Verbalization Categorization Strategies</i> on the Subtle Patterns that Indicate Usability Problems.....	169
9.2.4	Uncovering Subtle Patterns in <i>Other Types of Think-Aloud Protocols</i>	170
9.2.5	Discovering the <i>Subtle Verbalization & Speech Patterns</i> that are indicative of <i>the Severity of Usability Problems</i>	171
9.2.6	Creating a <i>Fully Automatic Pipeline</i> to Detect Usability Problem Encounters...	171
9.2.7	Leveraging the <i>Wisdom of the Crowds</i> to Detect Usability Problems.....	172
9.2.8	Uncovering Subtle Patterns in <i>Other Modalities (e.g., Gaze, Facial Expressions, Body Language, and Physiological Measures)</i> that Are Indicative of Users' <i>Negative and Positive</i> Experiences.....	173
9.2.9	Building ML Models that Leverage <i>Bigger</i> Dataset and Can Detect <i>Product- specific</i> and <i>User-specific</i> Problems <i>More Effectively</i>	174
9.2.10	Forging a Symbiosis Relationship between UX Practitioners and the ML/AI to better <i>Identify</i> and <i>Interpret</i> Usability Problems.....	175
9.2.11	Designing <i>Real-time Intelligent</i> Systems to Assist UX Practitioners to <i>Conduct</i> Usability Test Sessions <i>More Strategically</i>	178
References	181

List of Tables

Table 1. Correlation analysis between responders' profile information and their practices of conducting think-aloud sessions (* indicates significance).....	34
Table 2. Tasks that participants worked on the two products.....	64
Table 3. The counter-balancing scheme (p1-p8 denote the participants' IDs in the think-aloud data collection).....	65
Table 4. The frequency and percentage of the audio segments labeled with each verbalization category.....	69
Table 5. Precision, recall, and F-measure of each category in identifying problems for each test device.....	71
Table 6. Precision, recall, and F-measure of the verbalization category pairs in identifying problems.....	72
Table 7 Tasks for the two websites used in think-aloud data collection.....	83
Table 8 A balanced Latin Square design for every four evaluators.....	85
Table 9. The percentage of audio segments labeled with each verbalization category for each testing object.....	89
Table 10. The percentage of audio segments labeled with each verbalization category for physical and digital products respectively.....	91
Table 11. Precision, recall, and F-measure of each verbalization category in identifying problems for physical devices vs. digital websites.....	92
Table 12. Precision, recall, and F-measure of each verbalization category pair in identifying problems for physical devices vs. digital websites.....	93
Table 13. Precision, recall, and F-measure of each verbalization category in identifying problems when evaluators had access to the audio or video modality of the sessions.....	94

Table 14. Precision, recall, and F-measure of each verbalization category pair in identifying problems when evaluators had access to the audio or video modality of the sessions.	94
Table 15. Precision, recall, and F-measure of each verbalization category in identified problems when evaluators worked with or without visualization.	95
Table 16. Precision, recall, and F-measure of each verbalization category pair in identified problems when evaluators worked with or without visualization.	95
Table 17. Verbalization category proportion.	102
Table 18. the average any-two agreement between evaluators.	103
Table 19. The accuracy, precision, recall, and F1-score of the two-class (i.e., Observation and non-Observation) and four-class (i.e., Reading, Procedure, Observation, and Explanation) category classifiers.	124
Table 20. The performance of the ML models when trained with different sets of features.	131
Table 21. The number of problems identified by evaluators ($\mu(\sigma)$)	142
Table 22. The agreement and disagreement of identified problems between evaluators and ML.  : problems that evaluators and ML agreed;  : problems that only evaluators identified;  : problems that only ML identified. Results are shown as ($\mu(\sigma)$).	143
Table 23. The number of times for pauses and rewinds ($\mu(\sigma)$).	144
Table 24. The frequency of evaluators' session reviewing strategies based on passes.	147

List of Figures

Figure 1. Explicitly asking participants to explain their behavior during think-aloud sessions may change their thought processes for completing the task. The states in the bottom line are the unaltered states while the states in the top line are extra states that are triggered by external requirements [31].	16
Figure 2. Respondents' years of work experience in HCI/UX/usability testing fields.	26
Figure 3. The distribution of the size of the companies that respondents worked in.	27
Figure 4. The most frequently used methods for detecting usability problems.	27
Figure 5. The frequency of using concurrent-think-aloud protocols and retrospective think-aloud protocols among the respondents.	29
Figure 6. The testing environments in which UX practitioners use think-aloud protocols.	30
Figure 7. The types of tasks that UX practitioners ask their participants to work on during think-aloud sessions.	30
Figure 8. The frequency of conducting practice sessions before actual think-aloud sessions.	31
Figure 9. The content that UX practitioners ask their participants to verbalize in addition to the content that comes naturally into the participants' mind.	32
Figure 10. The percentages of different combinations of the content that respondents ask their participants to verbalize.	32
Figure 11. How the frequency with which respondents prompted their participants during think-aloud sessions had changed compared to when they just started their UX career.	33
Figure 12. The activities that UX practitioners perform when analyzing think-aloud sessions.	35
Figure 13. The types of information that help locate usability problems.	35
Figure 14. The types of information that UX practitioners seek in users' verbalizations.	36

Figure 15. The ways in which UX practitioners deliver their analysis results. 37

Figure 16. Participation in three types of data analysis activities: writing an informal usability test report; writing formal usability test report; having a data analysis discussion meeting. 38

Figure 17. The tool for evaluators to segment audio recordings of think-aloud sessions and provide category labels for each segment. The colored bar at the bottom shows the category labels that are already assigned by the evaluator. 50

Figure 18. The tool for evaluators to analyze a recorded think-aloud session. It visualizes the transcript of the session on the left panel. Seven audio features are represented as charts on the right panel. Highlighting any part of a chart will highlight the corresponding transcript on the left panel. The bottom part of the tool allows an evaluator to log problems and select the verbalization and speech features that they used to identify the problems. 52

Figure 19. Callout views of three parts of the analysis tool: seven feature panels (top); highlighted word of the current timestamp (middle); problem description area (bottom). 53

Figure 20. The visualization of the silence (the colored part of the timeline) before and after selecting a silence length filter. 54

Figure 21. An evaluator was analyzing a recorded think-aloud session during the study. 55

Figure 22. The verbalization categories for a think-aloud recording and the problems identified by an evaluator. 57

Figure 23. The number of times that each feature that evaluators used to identify problems. 58

Figure 24. (left) clinches on the speech rate graph; (right) irregular patterns appeared in both loudness and pitch graphs. 61

Figure 25. The updated tool for evaluators to analyze recorded think-aloud sessions. It visualizes the transcript of the recording on the left panel (a), one line per audio segment. The seven features (i.e., category, silence, verbal fillers, sentiment, speech rate, loudness, pitch) are represented as time-synchronized charts on the right panel (b). Selecting any part of a chart

highlights the corresponding transcript on the left panel. The bottom of the tool (c) allows for describing usability problems and the features used to identify the problems.	68
Figure 26. Precision, recall, and F-measure of each verbalization category in identifying problems.....	71
Figure 27. The number of times that each verbalization and speech feature was used by evaluators for finding usability problems.	73
Figure 28. The think-aloud analysis tool's interfaces for the <i>Audio only</i> condition	86
Figure 29. The think-aloud analysis tool's interface for the <i>Video only</i> condition.....	86
Figure 30. The think-aloud analysis tool's interface for the <i>Audio + Visualization</i> condition. ...	87
Figure 31. The think-aloud analysis tool's interface for the <i>Video + Visualization</i> condition. The left two columns are the same as the <i>Audio + Visualization</i> condition.....	87
Figure 32. Precision, recall, and F-measure of each verbalization category in identifying problems.....	90
Figure 33. Precision, recall, and F-measure of each verbalization category pair in identifying problems.....	91
Figure 34. Most frequently verbalized words when users encountered problems.....	96
Figure 35. Precision, recall, and F-measure of each speech feature in identifying problems.	98
Figure 36. The average F1-score of the four types of ML models trained using <i>the transcript feature only</i> as the input or using <i>the transcript and one additional verbalization & speech feature (i.e., negation, category, question, sentiment, pitch, and speech rate) together</i> as the input and evaluated the models using 10-fold cross-validation on the entire dataset.....	113
Figure 37. The precision, recall, and F1-score of the four types of ML models trained using the <i>transcript</i> feature only as the input (the left half of the figure) and using the <i>transcript + all verbalization & speech features</i> together as the input (the right half of the figure) and evaluated using 10-fold cross-validation on the entire dataset.	114

Figure 38. The average precision, recall, and F1-score of the four types of ML models trained using only the *verbalization & speech features* as the input and evaluated using 10-fold cross-validation on the entire dataset. 115

Figure 39. The precision, recall, and F1-score of the SVM models trained with each verbalization or speech feature respectively and together and evaluated using the 10-fold cross validation. . 116

Figure 40. The average precision, recall, and F-1 score of the SVM model trained on any seven users’ data using the *transcript (i.e., TF-IDF) + all the verbalization & speech features together* as the input and evaluated on the rest one user’s data for each *product* respectively (i.e., *leave-one-user-out* scheme). 117

Figure 41. The average precision, recall, and F-1 score of the SVM model trained on any three products’ data using *the transcript (i.e., TF-IDF) + all verbalization & speech features together* as input and evaluated on the rest one product’s data for each *user* respectively (i.e., *leave-one-product-out* scheme). 118

Figure 42. The initial way of visualizing verbalization and speech features. 133

Figure 43. VisTA: a visual analytics tool that allows UX practitioners to analyze recorded think-aloud sessions with the help of machine intelligence to detect usability problems. 134

Figure 44. The *problem timeline* of VisTA. The problem timeline highlights all the segments (i.e., c, d, and e in the chart) that have the same features as the currently paused timestamp (c), which allows evaluators to examine where in the session the same features appeared and how those areas align with the ML-inferred problems. 135

Figure 45. The feature timeline that shows the ML’s input features in a short time window around the current time in the video. The current time in the video is marked as the red vertical line in the center. 136

Figure 46. VisTASimple UI. Compared to VisTA, VisTASimple only presents the problem timeline without showing the ML’s input features or providing the feature filtering function. . 138

Figure 47. Baseline UI. Baseline has the same video reviewing and problem annotation functions as the VisTA and VisTASimple but it has no ML-related features. 139

Figure 48. Typical playback behaviours (x-axis: session time; y-axis: reviewed video time).. 146

Figure 49. The *problem timeline* with evaluator-identified problems visualized on top of it. Both the ML-inferred problems and the evaluator identified problems can act as "anchors" to facilitate the re-visitation. 155

Chapter 1 Introduction

Think-aloud protocols were first developed by Ericsson and Simon [32] to study the human thought processes in psychology and later introduced into HCI by Lewis to uncover usability problems with user interfaces in 1982 [59]. While thinking aloud, participants verbalize their thoughts when working on a task simultaneously; this enables evaluators to learn about problems encountered by potential users and gain insights that cannot be easily obtained from mere observations [24]. This type of thinking aloud is called concurrent thinking aloud and is the focus on this dissertation.

Despite the value of conducting think-aloud sessions, analyzing think-aloud sessions is often arduous. Although it is possible to have a usability evaluator observe and analyze think-aloud sessions on the fly, it is common practice to audio or video record think-aloud sessions for later analysis. A recent study found that think-aloud sessions were video recorded by 73% of the usability practitioners and reviewed by half of the usability practitioners [34]. The analysis process typically involves listening to or watching the session recordings, transcribing the session recordings, and reviewing the transcripts, and scrutinizing users' verbalizations among other information to pinpoint moments when users were experiencing problems [53,64].

One way to facilitate the analysis process is to have a usability evaluator observe the test session and take observation notes. However, this approach relies, to a large extent, on the evaluator's ability to catch the important moments and note down those moments on the fly. Moreover, because notes written on the fly tend to be brief and lack of contexts, the usability evaluator often needs to review the session recordings to pinpoint moments when the notes were taken to better understand their meaning and contexts.

Alternatively, another way to facilitate the analysis process is to have a team of usability evaluators observe a think-aloud test session, take notes and discuss their observations and insights afterward. This approach has a better chance of capturing the important moments of the session with different perspectives from multiple evaluators. However, it requires more than one usability evaluator to

collaborate and work together. In practice, few UX practitioners, however, have such an opportunity to analyze the same usability test session with their colleagues [34].

Furthermore, UX practitioners often work under time pressure and tend to perform quick, rather than rigorous, analysis. As a result, previous research has argued for developing methods to support fast-paced analysis of recorded think-aloud sessions [34,76].

Another motivating factor is agile usability evaluation, which has become increasingly popular over the past decade. Like agile software development, agile usability evaluation emphasizes rapid iterations, in this case conducting and analyzing more rounds of usability tests in different product development cycles so that more feedback learned from the usability tests can be incorporated into the development promptly. This further enhances the need for effective face-paced methods for analyzing recorded think-aloud sessions.

Previous research has shown that subtle patterns in how we interact with other people reveal our attitudes toward them [83]. For example, expert poker players can estimate whether their opponents are bluffing or not by reading the subtle signals in what they say and how they say it among other subtle honest signals. Inspired by this idea, I hypothesize that *subtle patterns in what users verbalize and how they verbalize it during think-aloud sessions can reveal their attitudes toward the test product (e.g., the problems that they encounter while using the product)*. The intuition for the hypothesis is that what people say and how they say it can reveal their cognitive load [24,30], emotions (e.g., excitement, frustration) [51,83,88], and the level of confidence (e.g., [38,57,83]).

If subtle patterns *in what and how users verbalize* are indeed indicative of the usability problems that they encounter, then there might be an opportunity to leverage these patterns to design computational methods and interactive tools to assist UX practitioners with performing analysis effectively.

1.1 Research Objective

In this dissertation, I first surveyed UX practitioners who work in different fields around the world to understand the practices and challenges around the use of think-aloud protocols. Informed by the findings, I then designed and conducted a series of three studies to understand whether and how *subtle patterns in what and how users verbalize during think-aloud sessions* are related to *when usability problems are encountered*. Informed by the understanding of the subtle verbalization and speech patterns that tend to occur when users encounter usability problems through these studies, I then designed and evaluated computational methods that leverage these subtle patterns and the machine learning (ML) algorithms to detect the encounters of usability problems in recorded think-aloud sessions. To help UX evaluators leverage ML-inferred usability problem encounters, I iteratively designed an interactive intelligent visual analytics tool that integrates machine intelligence with visualization to help UX evaluators identify when usability problems were encountered in a recorded think-aloud session. I further designed and conducted a controlled lab study, which demonstrates that UX practitioners can use ML-inferred problem encounters to analyze think-aloud sessions more effectively. Based on the lessons learned from this dissertation work, I present my *thesis statement* as follows:

Subtle verbalization and speech patterns tend to occur when users encounter problems in concurrent think-aloud sessions; these subtle patterns can be used to automatically detect usability problem encounters, which can be used by UX practitioners as overviews, aids, reminders, anchors, and guides to identify usability problems more effectively.

I have validated the thesis statement by answering the following four research questions (RQs):

- RQ1: What are the current practices and challenges that UX practitioners have when using think-aloud protocols?

To answer this question, I designed and conducted an international survey study to understand how UX professionals around the world currently use think-aloud protocols. Based on the results of 197 responses from UX professionals in six continents, I found that 86% of the respondents used

think-aloud methods when conducting usability tests. Additionally, analyzing think-aloud sessions is arduous but manual analysis approaches do not scale well with large quantities of usability tests.

- RQ2: What are the subtle patterns in *what* participants verbalize (i.e., verbalizations) and *how* they verbalize (i.e., speech features) that tend to occur when they encounter problems in think-aloud sessions?

To answer this question, I designed and conducted three controlled lab studies, each addressing the limitations of the previous one. The first two studies identified and validated the subtle verbalization and speech patterns that are related to the usability problems that think-aloud participants experienced when using different sets of products. I further conducted a third study to show that these verbalization and speech patterns that are indicative of usability problems are not affected by three external factors, which are the type of test products (i.e., physical products vs. digital products), the access to different modalities of session recordings (i.e., audio vs. video), and the access to the visualization of verbalization and speech features (i.e., presence vs. absence) during evaluators' analyses.

- RQ3: Can the subtle verbalization and speech patterns be used to detect the encounters of usability problems automatically?

To answer this question, I applied natural language processing (NLP) and machine learning (ML) methods to build models to detect usability problem encounters automatically. To understand whether and how the subtle verbalization and speech patterns help to detect the encounter of usability problems, I compared the effectiveness of different verbalization and speech features with commonly used text-based features.

- RQ4: Can UX practitioners use ML-inferred usability problem encounters to help them with their analysis?

To answer this question, I first iteratively designed an intelligent interactive visual analytics tool that presents ML-inferred usability problem encounters as a series of “spikes” on a timeline along

with other review and annotation functions to assist UX practitioners with their analysis. I then conducted a controlled between-subjects user study with three experimental conditions to better understand how UX practitioners would leverage and perceive ML-inferred problem encounters and ML's input features in their analysis. The results showed that UX practitioners can identify usability problems more effectively when presented with ML-inferred problem encounters than without these predictions.

1.2 Contributions

By answering the four RQs, this dissertation makes the following four contributions:

- An understanding of the current practices and challenges around the use of think-aloud protocols in industry. This was accomplished through a survey study which confirms the wide adoption of the method in industry across different fields and reveals the challenges associated with the analysis of think-aloud sessions among UX practitioners in different geographical regions and industrial fields;
- The identification and validation of subtle verbalization and speech patterns that are indicative of usability problems in concurrent think-aloud sessions. Specifically, when users encounter usability problems, their verbalizations tend to include the *Observation* category, negative sentiments, questions, negations, verbal fillers, and abnormally high or low pitches, or low speech rates;
- A demonstration of the effectiveness of these subtle verbalization and speech patterns in detecting usability problem encounters automatically. Specifically, the subtle verbalization and speech features that tend to occur when users encounter problems are informative and complementary to each other and can be used together to build the ML model that detects usability problems encounters with 0.75 F-measure; furthermore, results also show that effective user-dependent and product-dependent ML models can also be built to detect usability problem encounters.

- A demonstration of the effectiveness of ML-inferred problem encounters in helping UX practitioners identify usability problems more effectively through the design and evaluation of an intelligent visual analytics tool, VisTA, which presents the ML-inferred problem encounters as a series of “spikes” on a timeline and the ML model’s input features within a short time window among other functions to assist UX practitioners with analyzing recorded think-aloud sessions. Specifically, the results of the controlled three-session between-subjects study show that UX evaluators identified more problems when assisted with the machine intelligence than without. The results further reveal insights into how ML-inferred problem encounters and the ML’s input features were used by the UX evaluators to assist their analysis.

1.3 Dissertation Outline

Next, I present an overview of each of the remaining chapters and how each chapter ties back to the four RQs and contributions:

- In Chapter 2, I review the background and related work about the think-aloud protocols, verbalization categorization, and the use of speech features. The chapter provides the theoretical foundations for conducting think-aloud sessions and motivations for RQs.
- In Chapter 3, I present the design and results of a survey study, which aims to understand how think-aloud protocols are currently used by UX practitioners in different industrial fields around the world. Findings from the survey study point out the value of analyzing thinking-aloud sessions in locating usability problems and the need for more efficient data analysis methods. Chapter 3 answers *RQ1* and provides *the first contribution*.
- In Chapter 4, I present the first study (Study 1) to explore the subtle patterns in users’ verbalization and speech features that are indicative of usability problems in concurrent think-aloud sessions.

- In Chapter 5, I present the second study—the validation study (Study 2)—to address the limitation of the Study 1 and validate its findings with more participants and a different set of test devices.
- In Chapter 6, I present the third study—the generalization study (Study 3)— to assess the generalizability of the findings identified from the first two studies. Specifically, I demonstrate that the subtle verbalization and speech patterns that are indicative of usability problems are robust to the types of products (i.e., physical devices and digital systems), and the recording modality (i.e., video vs. audio) and the visualization of verbalizations (i.e., presence or absence) that UX evaluators had access to. Chapters 4, 5, and 6 answer *RQ2* and provide the *second contribution*.
- In Chapter 7, I present the details of the design and evaluation of the computational methods that leverage the subtle verbalization and speech patterns to detect the encounters of usability problems by harnessing the power of natural language processing and machine learning. I report a series of evaluations on a range of ML models trained on different input features and demonstrate the effectiveness of each verbalization and speech features in improving the detection of usability problem encounters. Chapter 7 answers *RQ3* and provides the *third contribution*.
- In Chapter 8, I present the design, implementation, and evaluation of the think-aloud analysis tool, VisTA, which visualizes ML-inferred usability problem encounters and ML’s input features among other functions to assist usability evaluators with identifying usability problems more efficiently. Chapter 8 answers *RQ4* and provides the *fourth contribution*
- In Chapter 9, I conclude the dissertation by reiterating the contributions, discussing the limitations and potential future research directions.

Chapter 2 Background and Related Work

In this chapter, I present the background and related work around think-aloud protocols, users' verbalizations during think-aloud sessions, and speech features that are relevant to this dissertation research.

2.1 Think-Aloud Protocols

2.1.1 History

Think-aloud protocols were initially developed in psychology to study human thought processes that cannot be observed via people's actions alone [32]. Ericsson and Simon introduced and developed the theoretical framework for two types of think-aloud protocols: *concurrent think-aloud*, in which participants verbalize their thoughts while working on a task and *retrospective think-aloud*, in which participants verbalize their thoughts after completing a task. Retrospective think-aloud is typically conducted by having participants observe the video recording of their study sessions after they completed them.

Think-aloud protocols were later introduced into the human-computer interaction (HCI) field by Lewis in the 1980s to study usability problems that users encounter while using a computer system [59]. Nowadays, think-aloud protocols are commonly included in textbooks as a key method to identify usability problems [26,70,86,87]. In practice, think-aloud protocols are considered to be the "gold standard" for usability evaluation [47] and the single most valuable usability engineering method [70].

2.1.2 Concurrent vs. Retrospective Think-Aloud Protocols

Previous research has compared *concurrent* think-aloud protocols with *retrospective* think-aloud protocols and found no difference in terms of task performance [77] or the total number of problems discovered [49]. Also, *concurrent* think-aloud protocols are considered to be more efficient, easier to perform and moderate [2], reduce biases arising from post-task rationalization [49,64], and have been shown to have a negligible influence on participants' behavior [41].

In this dissertation, I focused on *concurrent think-aloud protocols*. For brevity, I will refer concurrent think-aloud protocols as think-aloud protocols in the rest of the dissertation unless mentioned otherwise.

2.1.3 Comparison to Other Usability Evaluation Methods

Usability evaluation methods are typically inspection-based methods or user-based methods. Inspection-based methods uncover potential usability problems by having expert evaluators inspect user interfaces with a set of guidelines or questions [74]. Common inspection techniques include heuristic evaluation and cognitive walkthrough.

Heuristic evaluation is an inspection method that is carried out by experts to discover interface design issues by using a list of heuristics [36,72]. Compared to the heuristic evaluation, think-aloud protocols tend to identify more concrete usability problems related to the tasks used in the evaluation. Yen et al. used a web-based communication tool for nurse scheduling as the example product to compare the concurrent think-aloud protocol with heuristic evaluation for identifying usability issues [105]. In their study, five HCI experts were used to perform a heuristic evaluation using Nielsen's ten heuristics [72] on four tasks, and a group of end-users of the tool performed thinking aloud on a set of tasks. They found that experts revealed more general interface problems but end-users who used the think-aloud protocol uncovered more concrete interface issues related to the tasks.

Cognitive walkthroughs are designed to understand whether a new user can carry out tasks with a system. The method begins by defining the task or a set of tasks that the user would be expected to carry out and divides the task or the tasks into steps. Evaluators try to perform the steps and ask themselves a set of questions [98], whose answers would be used to identify usability problems. The cognitive walkthrough method can be used by individuals with limited formal training in user interface evaluation.

Compared to cognitive walkthroughs, the think-aloud protocol tends to identify more problems at all levels of severity. Beer et al. compared these two methods by asking users to work on tasks on

an air ticket purchasing website and discovered that while the cognitive walkthrough method found the most severe usability problems, the think-aloud protocol found all types of usability problems with all levels of severity [6]. Similarly, Karat et al. also found that empirical testing using the think-aloud protocol identified more problems than both the individual walkthrough and the team walkthrough [52].

2.1.4 Validity Confirmation with Other Streams of Data

A recent study leveraged functional magnetic resonance imaging (fMRI) and the think-aloud method to study whether the think-aloud protocol is a valid measure of thinking [27]. Seventeen internal medicine physicians first underwent formal think-aloud training. Next, they answered validated multiple-choice questions in an fMRI scanner while both answering (thinking) and thinking aloud about the questions. Results show that the same brain regions were activated during answering and thinking aloud and these regions were significantly more active during answering than thinking aloud. These findings add evidence to the notion that the think-aloud protocol is a useful operational measure of thinking.

2.2 Conducting Think-Aloud Sessions

2.2.1 Guidelines for Conducting Think-Aloud Sessions

To encourage participants to verbalize their authentic thought processes, Ericsson and Simon proposed three guidelines for practitioners to conduct think-aloud sessions [32]. The three guidelines are: 1) *keep the interaction to the minimal* (i.e., only remind participants to think aloud if they fall into silence for a period of time); 2) *use neutral instructions* (i.e., instructions should not request any specific type of content); and 3) *allow participants to practice thinking aloud*. These guidelines were followed by numerous studies (e.g., [3,24,30,41,65,109,110]). A think-aloud protocol that complies with the three guidelines is referred to as the **classic** think-aloud protocol.

The first guideline aims to minimize the interaction between the experimenter and the participant. Ericsson and Simon recommend that the evaluator (i.e., experimenter) use the token “keep talking”

as it is non-directive and does not require an answer to the evaluator. This would keep the participants focusing on the tasks. The second guideline recommends the evaluator ask their participants simply report anything that comes into the mind without requiring the participants to verbalize a specific type of content. The third guideline encourages the participants to get used to verbalizing their thought processes via practice. Having participants practice thinking aloud can be helpful to eliminate silence due to misunderstandings of the instructions to think aloud and are recommended by prior research [16,32].

2.2.2 Relaxed Think-Aloud Protocols

Relaxed think-aloud protocols comply with the general procedure of the *classic* thinking aloud protocol but violate the Ericsson and Simon's guidelines about the use of reminders and instructions. For example, when using relaxed think-aloud protocols, the evaluator may probe participants by asking questions instead of simply reminding them to keep talking after long period of silence, or use directed instructions to request a particular type of content (e.g., asking participants for explanations and comments) instead of neutral instructions (i.e., asking participants to verbalize whatever naturally comes into their mind) as recommended. Boren and Ramey were the first to document this gap between the practice and the theory of thinking aloud protocols [10], and they proposed one type of relaxed think-aloud protocol—the speech-communication (SC) protocol. Relaxed think-aloud protocols encourage users to verbalize their thoughts (e.g., [40,110]) and are often used by UX practitioners [64].

2.3 Classic vs. Relaxed Think-Aloud Protocols

2.3.1 The Effect of Interaction Between the Evaluator and the Participant

Interactions between the evaluator and the participant in a think-aloud usability test session, such as probing the participant with questions, have shown to affect the participants' performance [2,42,79,109]. Zhao and McDonald compared the classic think-aloud protocol with a relaxed think-aloud protocol, in which the evaluator interacted with participants to gather explanations about their interactions and experiences [109]. Results showed that the percentage of relevant

utterances was not much higher in the relaxed think-aloud condition than the classic think-aloud condition, but participants in the relaxed think-aloud condition felt the evaluators' interventions (i.e., the interaction between the evaluator and the participants) were distracting.

Compared to Zhao and McDonald's study, Olmsted-Hawala et al. added one more relaxed think-aloud protocol (i.e., the speech-communication protocol) and compared three types of think-aloud protocols (i.e., the classic protocol, the speech-communication protocol, and the coaching protocol with active intervention) to the silence condition using website related tasks [80]. They found that there was no difference in the task completion time between conditions, but the coaching protocol condition significantly affected users' performance as well as subjective satisfaction compared to the other two conditions.

In a recent study, Alhadreti and Mayhew investigated the use of three types of think-aloud protocols in website usability testing: the classic think-aloud, the speech-communication (SC) relaxed think-aloud, and the active intervention relaxed think-aloud [2]. When using the SC relaxed think-aloud protocol, the evaluator took on an active listener role and used tokens in the form of the affirmatory "Mm-hmm" with intonation or asking questions "and now...?" [10]. Results showed that the three types of protocols allowed the evaluator to identify a similar number of usability problems and types. However, the evaluator's active interventions in relaxed think-aloud protocols modified participants' behavior at the interface and negatively affected their feelings towards evaluation. Furthermore, relaxed think-aloud protocols required much greater investment in terms of evaluators' time. Thus, they recommended that it is safer and cheaper to follow the classic think-aloud protocol.

2.3.2 The Effect of Instructions for Requesting Verbalizations

Using *directed* instructions, the ones that require participants to verbalize specific types of content, can affect their task performance and alter their behavior. Wright and Converse compared the relaxed think-aloud protocol in which participants were asked to provide explanations with the silent condition and found that the task performance was significantly different between the relaxed think-aloud condition and the silence condition [99].

Hertzum *et al.* took a step further and added the classic think-aloud protocol into the comparison and investigated whether participants that think aloud concurrently in the classic or relaxed way behave differently compared to performing in silence [41]. In the relaxed thinking aloud condition, the experimenter explicitly *asked participants to report explanations and comments*. Results showed that compared to working in silence, classic thinking aloud has little or no effect on users' behavior except prolonging task completion time. However, relaxed thinking aloud affects users' behavior in many ways: participants took longer to solve tasks, performed more commands to navigate both within and between the pages of the testing websites, and experienced a higher mental workload.

In another study, McDonald and Petrie investigated whether the classic think-aloud protocol and the relaxed think-aloud with an explicit instruction lead to different task-solving performance compared to silent working [65]. Results showed again that the classic protocol had no impact on task performance compared to the silent working condition, but the relaxed think-aloud with an explicit instruction led to altered behavior (i.e., an increase in within-page and between-page navigation and scrolling activities). This finding further confirmed that of the Hertzum *et al.*'s study.

Fox *et al.* took a different approach by selecting 64 academic articles that used thinking aloud protocols and conducting a meta-analysis of 94 studies in these articles to compare concurrent think-aloud conditions with the silent condition. The analyzed studies were conducted between 1983 and 2009 and together involved 3462 participants. Through statistical analysis of these studies' data, Fox *et al.* further confirmed that: 1) asking participants to verbalize their thoughts during a task alone (i.e., the classic think aloud protocol) did not alter performance; 2) directing participants to provide explanations during a task (i.e., relaxed think-aloud protocols) changed their task performance [35].

2.3.3 Summary

The classic think-aloud protocol, in which the three guidelines are followed, is considered to be more efficient, easier to perform and moderate [2], avoids biases arising from post-task

rationalization [49,64], and has been shown to have no or negligible influence on participants' task performance or their behavior [10,41,65,80,109].

In contrast, when using relaxed concurrent think-aloud, in which evaluators actively probe or ask users questions [37,64] or use explicit instructions that request participants to verbalize a particular type of content, participants' behavior, task performance, and mental workload are often affected (e.g., [2,35,41,55,65,85,95,99]). As a result, in this dissertation, I followed *Ericsson and Simon's three guidelines to conduct classic think-aloud sessions* for all the studies to avoid potential threats to the validity of users' verbalizations.

2.4 Verbalizations in Think-Aloud Sessions

2.4.1 Three-Levels of Verbalizations

Verbalizations are the utterances that participants verbalize during the think-aloud sessions. These verbalizations are different from the utterances produced during inter-person communication in the sense that the verbalizations in think-aloud sessions are self-directed and have found to be more idiomatic and use more idiosyncratic referents than the utterances in the inter-person communication [32,97].

Ericsson and Simon categorized users' verbalizations during think-aloud sessions into three levels [32]. Level-1 (L1) verbalizations refer to information that is reproduced in the form in which it was acquired from the central processor without intermediate processes occurring between the subject's focus of attention to the information and its delivery [32]. In other words, L1 verbalizations are stored in verbal form.

Level-2 (L2) verbalizations occur when the thoughts being verbalized are in the subject's focus of attention but are stored in a non-verbal form. It involves explication of the thought processes, such as giving a thought a label or some other verbal reference before verbalizing it out. This level of verbalizations does not retrieve extra information beyond the one that is available in the focus of the subject's attention. However, it does require the subject to explicate the information that is encoded in a non-verbal format. Since explication or recoding the content into verbal format takes

time, a subject who is verbalizing at L2 level can take more time to complete the task than the one who does not verbalize. But this process does not change the structure of the subject's main thought processes [32].

Level-3 (L3) verbalizations require users to access their long-term memory, involving additional mental processing that may influence their focus of attention. For example, requesting an explanation from participants during think-aloud sessions, as illustrated in Figure 1, may change the participants' task performance and lead users to report L3 verbalizations [18,20]. When instructional procedures conform to the notion of L1 and L2 verbalizations, Ericsson and Simon, via analysis of large amounts of studies, asserted that there was no evidence that the verbalization process changes the course or structure of the thought processes [32]. In contrast, asking participants to report a *specific type of information*, such as explaining the problems encountered, or probe them with questions during the study session, can potentially cause them to filter, alter, or reorganize their thought processes to satisfy the requirements.

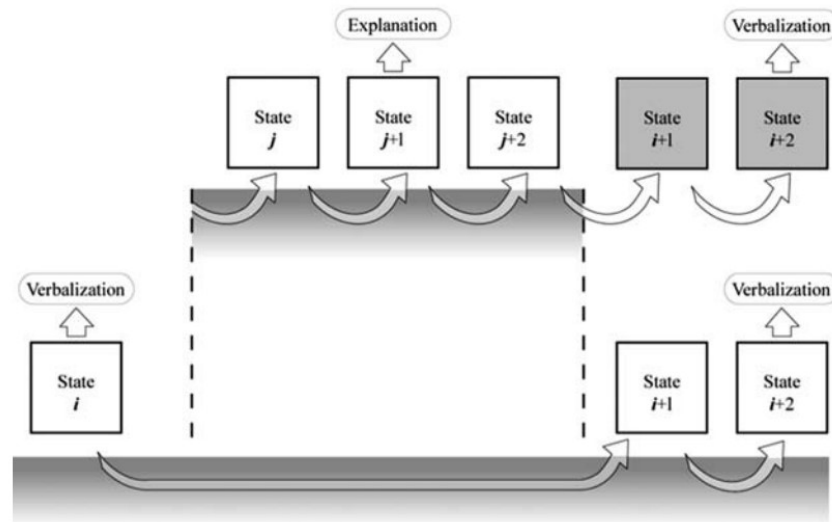


Figure 1. Explicitly asking participants to explain their behavior during think-aloud sessions may change their thought processes for completing the task. The states in the bottom line are the unaltered states while the states in the top line are extra states that are triggered by external requirements [31].

In the context of the usability studies, UX practitioners often care more about understanding users' experiences with the testing interface (i.e., L3 verbalizations) than users' mere action descriptions (i.e., L1 and L2 verbalizations) [10]. Therefore, L3 verbalizations, such as comments and explanations, can be informative for usability testing.

2.5 Verbalization Categorization

Although the guidelines for conducting the classic think-aloud sessions require the evaluator to not explicitly instruct participants to verbalize a specific type of content or to actively probe them with questions to explain their behaviors, the streams of participants' reported thought processes can still possibly contain explanations and comments of their actions in addition to their action descriptions. For example, early work from Bowers and Snyder found that most verbalizations during classic concurrent think-aloud sessions were descriptions of participants' onscreen behavior, but a portion of the verbalizations was about explanation and design [11]. Kraemer and Ummelen also found that their participants in the classic think-aloud condition verbalized comments (e.g., "I feel a little as if what I'm doing is hopeless") and explanations ("I am just doing something...because I don't really know where to look.") [55].

To better understand the verbalizations (i.e., utterances) that users verbalize during think-aloud sessions, Cooke conducted classic think-aloud sessions in which participants completed tasks on a website and recorded their verbalizations [24]. The verbalizations were scrutinized and segmented into small segments based on the pauses between utterances and the meaning of the utterances. Based on the analysis, Cooke proposed a coding scheme to categorize users' *verbalizations* into *four categories*: Reading, Procedure, Observation, and Explanation. The Reading category refers to when *the user reads words, phrases, or sentences directly from the product or its instruction manual*. The Procedure category refers to when *the user describes his/her current/future actions*. The Observation category refers to when *the user makes remarks about the product, its instruction manual or himself or herself*. Finally, the Explanation category refers to when *the user explains their motivation for their behavior*.

Elling *et al.* replicated Cooke's study procedure on three different websites [30]. In their analysis, they introduced two more categories (i.e., the Task-related category and verbal fillers category) to categorize their collected verbalizations better. The study results showed that 60% of the verbalizations fell into the *Observation*, *Explanation*, *Task-related*, and the *Verbal Fillers* categories, which demonstrated that verbalizations in classic think aloud sessions can provide information with an added value about the participants' processes and obstacles on the test interface.

Instead of using digital websites as test products as Cooke and Elling *et al.* did in their studies respectively [30], Hori *et al.* asked their participants to perform tasks on two physical products (i.e., a prototype of a touch-screen digital camera and a working product of photo album software) while thinking aloud. Their study revealed that users' verbalizations contained three of the four categories from Cooke's study [46] (i.e., Procedure, Observation, and Explanation).

Other studies extended Cooke's four categories into subcategories [40,110]. While Hertzum *et al.* broke down the Observation category further into four fine-grained sub-categories (i.e., system observation, redesign proposal, domain knowledge, and user experience) [40], Zhao *et al.* broke down the Observation category into three different sub-categories (i.e., expectation, positive experience, and negative experience) [110]. Similarly, Zhao and McDonald broke the Observation category into six subcategories (i.e., result evaluation, user experience, problem formulation, impact, recommendation, and task confusion). Additionally, they also broke the Explanation category into two subcategories (i.e., action explanation and causal explanation) [109]. This fine-grained verbalization categorization scheme was adopted by McDonald *et al.* for comparing users' utterances in concurrent and retrospective think-aloud usability test [66].

Inspired by prior works on categorizing verbalizations (i.e., utterances) in think-aloud usability test that followed Cooke's four categories, in this dissertation, I adopted Cooke's four categories to categorize users verbalizations and aimed to understand *how different verbalization categories link to usability problems* (e.g., which verbalization category users tend to verbalize when they encounter usability problems?).

2.6 Speech Features

Speech features, the characteristics of speech, can reveal speakers' feelings and mood [83]. The pitch of the user's voice may become higher when users are excited or surprised.

Additionally, speech features can also reveal speakers' confidence in conversations, which could, in turn, affect others' perception of their competence. Soman and Madan showed that the frequency and energy of interviewees' speech and their use of short words (e.g., OK?, yes!,) and interjections (e.g., uh, um) could be used to predict the outcomes of their job interviews [92]. Naim *et al.* further demonstrated that speech rate, pauses, speech coherence, and positive emotions conveyed via speech correlate positively with the overall interview performance and the likelihood of hiring [68].

Furthermore, speech features have also shown to be able to reveal speakers' cognitive load. When experiencing high cognitive load, users may compensate for the increased cognitive demand by directing more attention toward the task at hand, causing them to slow their speech down [38,91,94], fall into complete silence [32,84], decrease the volume of their voice [24,30] or use more verbal fillers (e.g., "um", "ah") [22,24,30]. Furthermore, Berthold and Jameson showed that users' spoken disfluency, speech rates, and pause rates vary when users experience different levels of cognitive load [7]. Yin *et al.* further demonstrated that the pitch and intensity of speech along with other acoustic features together could be used to build machine learning models to detect users' cognitive load automatically [107]. Additionally, the rate of pauses (i.e., silences between speech) was also shown to be effective to predict cognitive load [106].

Inspired by this line of research that explores the connection between users' speech features and the various aspects of their mental and emotional states, I hypothesize that users' speech features may exhibit certain patterns when they experience usability problems. Therefore, in this dissertation, I was motivated to understand *how speech features* (e.g., sentiment, silence, verbal fillers, speech rate, loudness, and pitch) *are linked to usability problems* (e.g., how do users tend to verbalize their thoughts when they experience usability problems?).

2.7 Think-Aloud Protocols Use in the Industry

McDonald *et al.* conducted a survey study in 2010 to understand the practices of think-aloud protocol use in academia and industry [64]. They found that: 1) think-aloud protocols were widely used in the industry; 2) 68% of their responders indicated that they used a general think-aloud protocol that was in line with Ericsson and Simon [32]; 3) while analyzing think-aloud sessions, they often had to review test notes, transcript and review session transcript, and review test videos.

Analyzing think-aloud sessions both rigorously and efficiently is a challenge. Nørgaard and Hornbæk observed 14 think-aloud sessions that were carried in seven companies located in Northern Europe to understand how the UX practitioners analyzed think-aloud sessions [76]. They found that systematic analysis was often not carried out and consequently encouraged the HCI/UX research community to develop methods to support fast-paced analysis and alleviate the time pressure that UX practitioners often face. The need for developing methods to support fast-paced analysis of think-aloud sessions was also echoed by Følstad *et al.*' study [34].

Inspired by the need for fast-paced analysis, in this dissertation, I aimed to understand whether subtle verbalization and speech features that are indicative of usability problems and whether these subtle patterns can be leveraged to speed up the analysis of thinking aloud sessions, perhaps via designing computational models and intelligent visualizations.

2.8 Summary

The synthesis of the literature indicates that the *classic concurrent think-aloud protocol* has little or no effect on users' task performance or their thought processes if Ericsson and Simon's three guidelines are followed. Violating any of these guidelines could introduce artificial changes to users' behavior and task performance. Therefore, in this dissertation, I adhere to these *three guidelines* to conduct the classic think-aloud sessions. Specifically, I include a *practice session* before actual test sessions for participants to practice and get used to thinking aloud while working on the tasks. I use the neutral instructions and only ask participants to say out the thoughts that naturally come into their mind without requesting any specific type of content. I keep the

interaction to the minimum by only reminding participants to “keep talking” when they fall into silence for a long period of time.

Previous research has shown that users’ verbalizations in think-aloud sessions can be categorized into four categories, each of which describes a different aspect of their behavior or thoughts. It is unclear, however, how these categories are related to usability problems. Do users tend to verbalize the content of one category over others when they encounter usability problems? Previous research has also shown that users’ speech features indicate their confidence levels, feelings, moods, and cognitive load. Inspired by this line of research, I am also motivated to explore how users’ speech features vary when they encounter usability problems. Specifically, do users tend to verbalize their thoughts with particular speech features when they encounter usability problems? These two questions together form *RQ2* of this dissertation: *What are the subtle patterns in what users verbalize (i.e., verbalizations) and how they verbalize it (i.e., speech features) that tend to occur when they encounter problems in think-aloud sessions?*

This literature review also found that think-aloud protocols are widely used in the industry and analyzing think-aloud sessions is labor-intensive. These findings, however, were based on survey or interview studies that were conducted either a long time ago or with a specific geographic region and a handful of participants. It is unclear whether these findings still hold today in industry. Therefore, before systematically answering *RQ2*, I first set out to understand whether the practices of using think-aloud protocols reported in the literature are still true today in different industrial fields around the world. To do so, I designed and conducted a *survey study* with UX practitioners around the world to understand *the current practices and challenges that UX practitioners encounter when conducting and analyzing think-aloud sessions*, which is *RQ1* of this dissertation. In the next chapter, I describe the detail of the survey study and its findings.

Chapter 3 Practices and Challenges of Using Think-Aloud Protocols in the Industry

Think-aloud protocols are one of the classic methods often taught in universities for training UX designers and researchers. Although previous research reported how these protocols were used in the industry, the findings were typically based on the practices of a small number of professionals in specific geographic regions [76,89] or on studies conducted years ago [64]. As UX practices continuously evolve to address new challenges emerging in the industry, it is important to understand the challenges faced by current UX practitioners around the world in a wide range of practical contexts when using think-aloud protocols. Such an understanding would inform the design of systems and methods that could potentially help UX practitioners better analyze think-aloud sessions. In sum, this chapter answers the first research question of this dissertation that has been introduced in Chapter 1:

RQ1: What are the current practices and challenges that UX practitioners have when using think-aloud protocols in the industry?

3.1 Introduction

Think-aloud protocols are often taught in UX courses to train professionals [26,70,86,87] and are considered as the “gold standard” for usability evaluation [47]. However, there has been little reported about how the protocols are used in practice. Previous research has examined the practices of using think-aloud protocols in local geographic regions. For example, Nørgaard and Hornbæk studied a small number of UX practitioners’ practices in Danish enterprises and offered insights on how they conducted and analyzed think-aloud sessions [76]. Similarly, Shi reported practices of and particular challenges in using think-aloud protocols [89]. In contrast, McDonald, Edwards, and Zhao conducted an international survey study to understand how think-aloud protocols were used in a broader scale and distributed the survey to UX professional and academic listservs [64]. However, as the survey was conducted eight years ago in 2011 and new UX testing software and tools have emerged over this period, the extent to which think-aloud protocols are currently being

used in industry is unclear. Moreover, recent research has also urged the community to learn more about the current UX practices in industry [62].

To better understand how think-aloud protocols are currently used in the industry, I designed and conducted a survey study. In total, 197 UX practitioners of different levels of experience, working in numerous industries around the world provided full responses to the survey. I present and discuss the key findings and implications of the survey study to inform UX practitioners about the practices and challenges surrounding the use of think-aloud protocols in the industry.

3.2 Goal

The goal of the study was to understand how think-aloud protocols are being used by UX professionals in different fields around the world. I chose surveying over other methods (e.g., interview, focus groups) because it allowed me to gather data from a broad range of UX practitioners located in different geographic regions who work in different industrial fields.

3.3 Participants

I contacted the organizers of local chapters of the User Experience Professional Association (UXPA), the largest organization of UX professionals around the world, to promote the survey study. I received support from the organizers of the UXPA's local chapters in Asia, Europe, and North America, who helped me to distribute the survey link to their listservs. I also promoted the survey link in a number of UX professional LinkedIn groups and other social media platforms. I conducted the survey study for about three months July-September 2018.

3.4 Survey Design

The survey was conducted as an online questionnaire using Google Form. The survey contained a list of multiple-choice and short-answer questions to understand whether and how UX professionals are currently using think-aloud protocols in addition to their basic profile information (i.e., the organization and the usability testing team that they work in and their current positions). No personally identifiable information was collected.

I was inspired by a prior survey study conducted in 2010 [64] but at the same time made important changes. The previous survey was distributed to UX practitioners working in both academia and the industry, which made it hard to isolate the use and practical impact of think-aloud protocols in the industry. Instead, this survey study was focused on the practices around the use of think-aloud protocols in the industry and thus was only distributed to UX practitioners in the industry. I also collected the participants' years of experience as a UX professional. Furthermore, as new tools and procedures for conducting usability test sessions have entered the market since 2010, such as the Agile-UX design [50], I wanted to see how the use of think-aloud protocols has evolved in light of the introduction of new practices.

3.5 Data Analysis

Answers to multiple-choice questions are quantitative data and were analyzed to identify the statistical trends in using think-aloud protocols. Answers to short-answer questions are qualitative data. Two researchers first independently analyzed the qualitative data using open coding and then discussed to resolve any conflicts. They then used affinity diagramming to identify common themes that emerged from the data.

3.6 Results

There were 197 valid responses to the survey from UX practitioners around the world. In this section, I report the aggregated information about the respondents' profile information and their practices of conducting and analyzing think-aloud usability tests.

3.6.1 Respondents' Profile

Location: In terms of the geographic locations, the majority of the survey respondents worked in North America 54% (n=125), followed by 19.3% Asia (n=38) and 14.7% Europe (n=29). Other respondents worked in Australia 1.5% (n=3), Africa 0.5% (n=1), and South America 0.5% (n=1).

Work Role: The majority of the respondents reported their current job title as a UX researcher (54%) or UX designer (36%). Others identified their job title as UX team lead (11%), UX manager

(8%) or design strategist (6%). Note that respondents were allowed to report more than one role that they had.

Work experience: The respondents had a wide distribution of work experience (Figure 2). 28% had *10 or more years* of work experience in HCI/UX/usability testing fields; 17% had *6-9 years* of work experience; 22% had *3-5 years* of work experience; 20% had *1-2 years* of working experience, and 13% had *less than one year's* work experience.

Companies or organizations: 81 respondents also reported the company that they worked in. There were 60 different companies reported. These included large enterprises and independent consultants, and covered a wide range of fields, such as IT (e.g., Google, Tencent), gaming companies (e.g., Ubisoft), banks and financial institutes (e.g., royal bank of Canada, PWC), telecommunication (e.g., T-Mobile, Telstra), health care (e.g., Blue Cross and Blue Shield, Klick), UX consulting (e.g., End to End User Research, Centralis), and software (e.g., Autodesk, SAS).

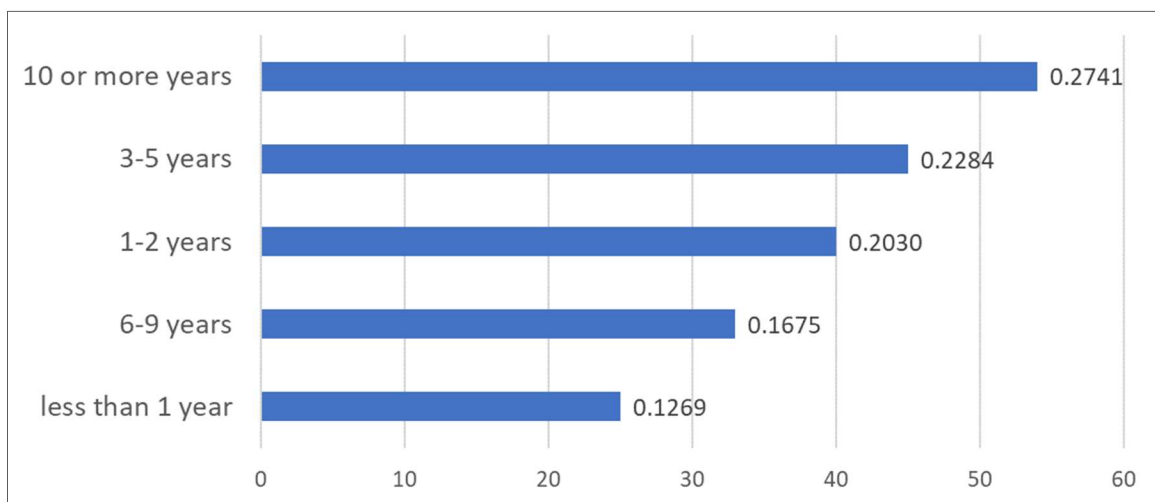


Figure 2. Respondents' years of work experience in HCI/UX/usability testing fields.

For the respondents who did not report their working companies, I asked them to report their company size information. For the ones that reported their companies, their sizes were searched

and found online. Figure 3 shows the distribution of the size of the companies that the respondents worked in.

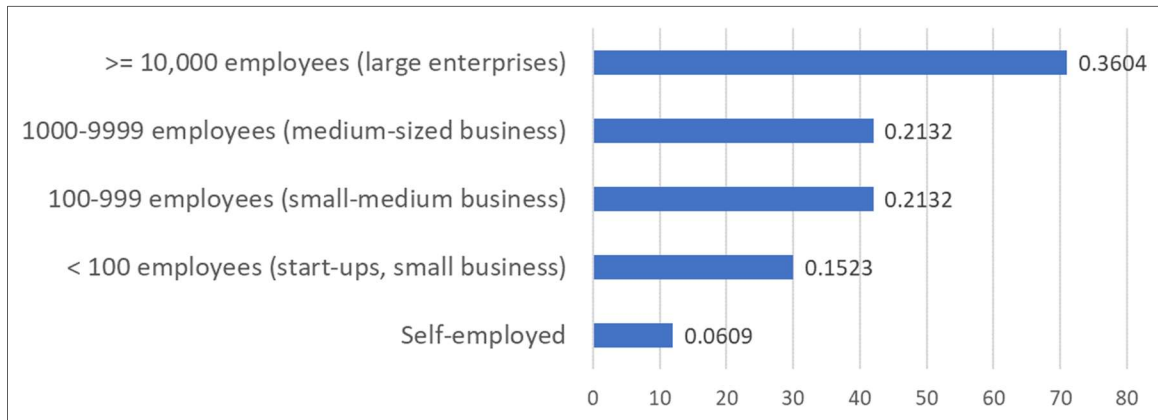


Figure 3. The distribution of the size of the companies that respondents worked in.

UX team size: Respondents worked in different sized UX teams: 1 (n=21), 2-5 (n=55), 6-10 (n=42), 11-15 (n=22), 16-20 (n=16), 20-30 (n=16), 30-50 (n=5), and >50 (n=20).

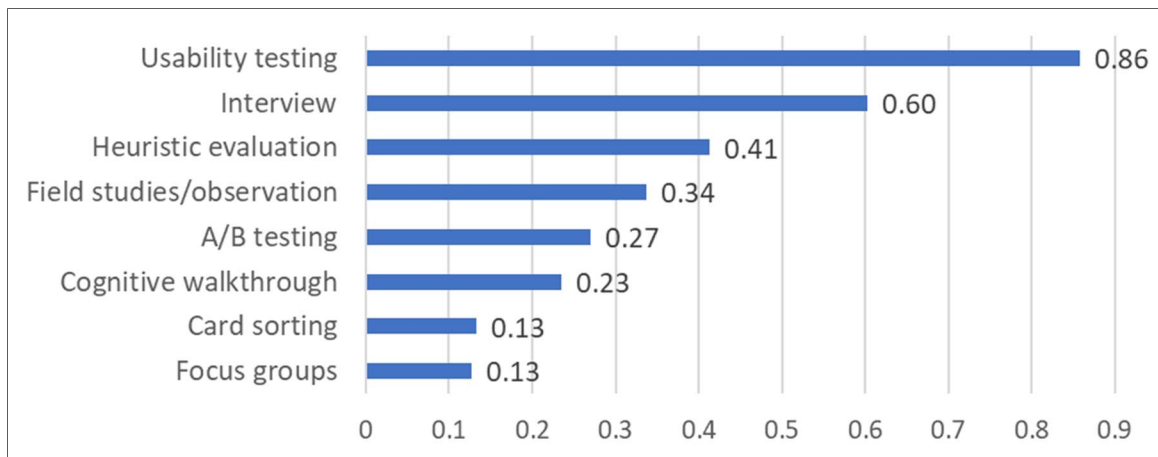


Figure 4. The most frequently used methods for detecting usability problems.

Methods for detecting usability problems: I asked respondents to provide their three most frequently used methods for detecting usability problems (Figure 4). The most frequently used

methods for detecting usability problems among the participants are as follows: usability testing (86%), interview (60%), heuristic evaluation (41%), field studies/observation (34%), A/B testing (27%), cognitive walkthrough (23%), card sorting (13%), and focus groups (13%).

3.6.2 General use of think-aloud protocols

Among the 197 respondents, 91% of them (n=179) reported that they had learned how to use think-aloud protocols and the remaining 9% (n=18) reported that they were unfamiliar with think-aloud protocols. For the 179 respondents who had learned think-aloud protocols, 49% of them (n=87) reported that they had learned the protocols in *university/college*, 36% (n=65) *at work*, and 15% (n = 26) from *UX online/offline bootcamps*.

General use and non-use of think-aloud protocols: When conducting usability tests, 86% of all the participants (n=169) reported that they used think-aloud protocols. In other words, 95% of the participants who had learned think-aloud protocols (169 out of 179) used them. I carried out the following analysis based on the responses of these 169 respondents who used think-aloud protocols because the remaining survey questions were about how UX practitioners used think-aloud protocols.

I also asked those respondents who had learned think-aloud protocols but did not use them (n=10) about their reasons for not using the protocols as an optional short-answer question and received 7 responses. The five reasons were as follows: 1) conducting think-aloud sessions is not part of their role (n=2); 2) their study subjects may not verbalize their thoughts easily (e.g., children) or unbiasedly (e.g., internal users) (n=2); 3) conducting think-aloud sessions takes too much time (n=1); 4) think-aloud protocols may distract their users (n=1); 5) there are alternative methods (n=1).

The frequency of use for two types of think-aloud protocols: There are two types of think-aloud protocols: *concurrent think-aloud protocols*, in which users verbalize their thoughts while working on tasks, and *retrospective think-aloud protocols*, in which users verbalize their thoughts only after they have completed the tasks (usually via watching their session recordings). I asked respondents

about their frequency of using *concurrent* and *retrospective* think-aloud protocols. Figure 5 shows the frequency of using concurrent and retrospective think-aloud protocols among the participants. Specifically, 61% of them (n=103) used the concurrent think-aloud protocols in *almost every usability tests* and 91% of them (n=154) used the concurrent think-aloud protocols in *at least half* of their usability tests. In contrast, only 21% of them (n=36) used the retrospective think-aloud protocols in *almost every usability tests* and the majority of them (61%, n=104) *almost never* or *only occasionally* (i.e., roughly a quarter of the tests) used the retrospective think-aloud protocols.

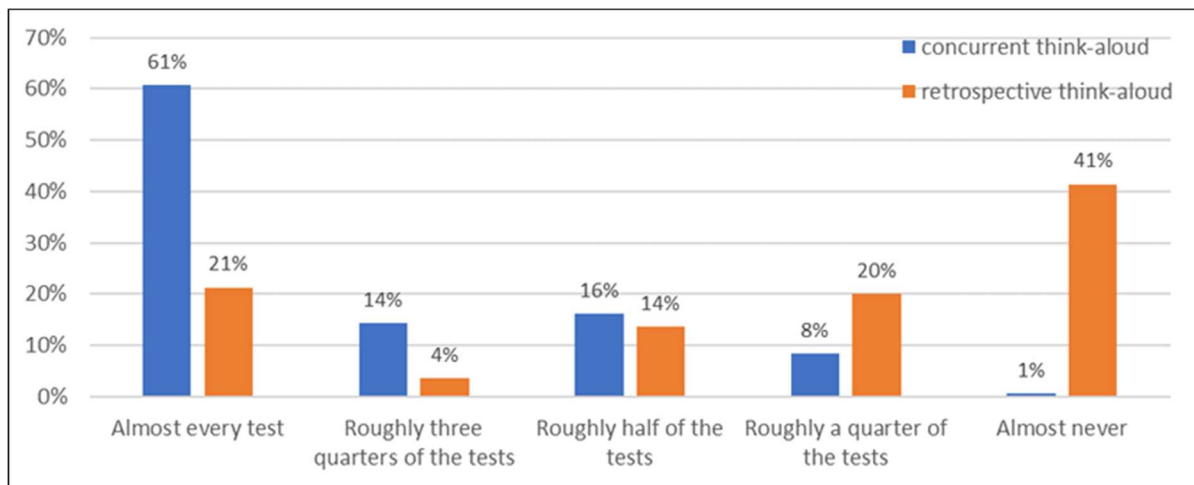


Figure 5. The frequency of using concurrent-think-aloud protocols and retrospective think-aloud protocols among the respondents.

Motivation: Respondents were asked about their motivation for using think-aloud protocols and found that 51% of the respondents (n=86) used the think-aloud protocols to both *inform the design* (e.g., *problem discovery*) and *to measure the performance* (e.g., *success rate*); 48% of them (n=81) *only* used the protocols *to inform the design* and only 1% of them (n=2) *only* used the protocols *to measure the performance*.

Testing environments: Figure 6 shows the testing environments in which the survey participants use think-aloud protocols. Specifically, 75% of the respondents (n=127) used the protocols in *controlled lab studies*, 72% of them (n=121) used the protocols in *remote usability testing*, and

48% of them (n=81) used the protocols in *field studies*. The total does not sum up to 100% because respondents can use the think-aloud protocols in more than one test environment.

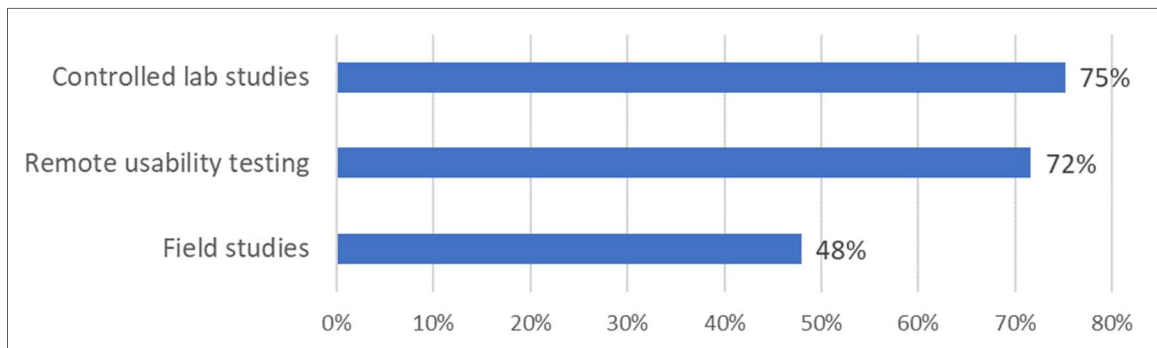


Figure 6. The testing environments in which UX practitioners use think-aloud protocols.

3.6.3 Conducting think-aloud sessions

Types of tasks for think-aloud sessions: Figure 7 shows the types of tasks that the respondents (i.e., UX practitioners) ask their participants to work on during think-aloud sessions. Specifically, 27% of them (n=46) *only* ask their participants to work on *tasks without instruction steps to follow* (e.g., navigating a website), while 12% of them (n=20) *only* ask their participants to work on *tasks with instruction steps to follow* (e.g., setting up a TV with its manual). In contrast, the majority of the respondents (61%, n=103) used *both* two types of tasks during think-aloud sessions.

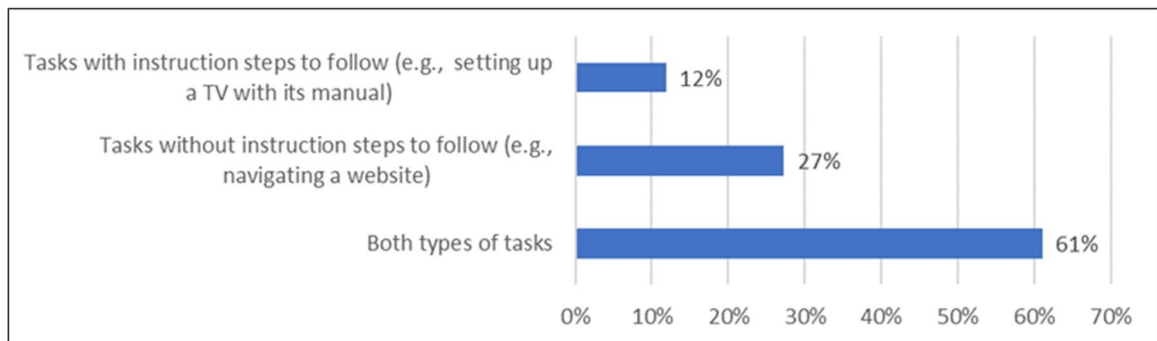


Figure 7. The types of tasks that UX practitioners ask their participants to work on during think-aloud sessions.

Practice sessions: Ericsson and Simon have suggested that practitioners should ask their participants to practice thinking aloud before conducting the actual think-aloud sessions (Ericsson & Simon, 1984). Figure 8 shows the frequency of conducting practice sessions before actual think-aloud sessions among the survey participants. Specifically, the majority of the respondents (61%, n=103) *almost never* conduct practice sessions, 7% (n=12) only do it *roughly a quarter* of the time, 6% (n=10) do it *roughly half* of the time, 2% (n=4) do it *roughly three-quarters* of the time, and 24% (n=40) do it *almost all* the time. The result shows that the majority of the UX practitioners seldom ask their participants to do a practice think-aloud session before conducting the actual think-aloud sessions.

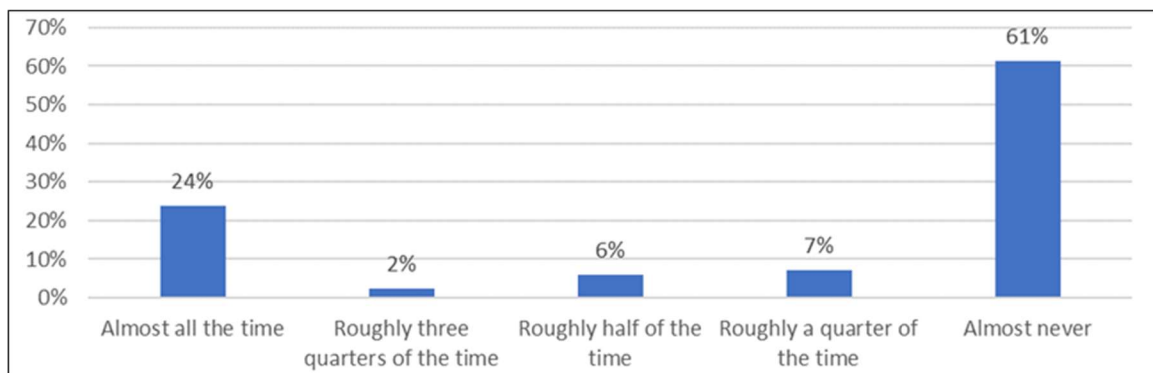


Figure 8. The frequency of conducting practice sessions before actual think-aloud sessions.

Instructions for requesting verbalizations: When using the classic protocol, UX practitioners are only recommended to ask their participants to say out loud everything that naturally comes into their mind. The survey asked the survey participants what else they ask participants to verbalize during think-aloud sessions. Figure 9 shows the results. Specifically, only 7% of the survey respondents (n=12) reported that they do not ask their participants to verbalize anything beyond what naturally comes into their mind. In contrast, 80% (n=136) mentioned that they also explicitly ask their participants to verbalize *their feelings*; 70% (n=119) explicitly ask their participants to verbalize *their feedback*; 55% (n=93) explicitly ask their participants to verbalize *their actions on the interface*, and 33% (n=55) explicitly ask their participants to verbalize their *design recommendations*.

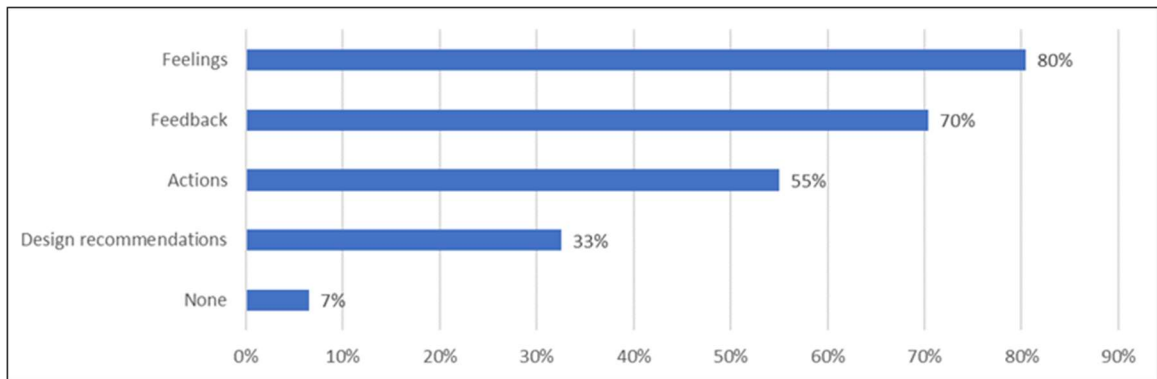


Figure 9. The content that UX practitioners ask their participants to verbalize in addition to the content that comes naturally into the participants' mind.

To better understand what types of content that respondents often request their participants to verbalize together, I further counted the number of occurrences of different combinations of content that they ask their participants to verbalize in addition to the thoughts that come naturally into the mind. Figure 10 shows the result.

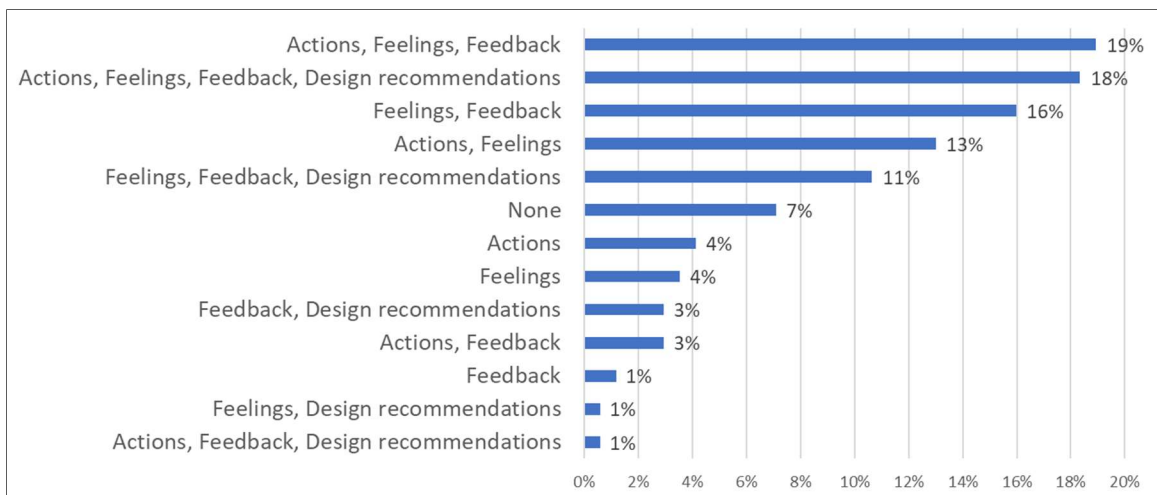


Figure 10. The percentages of different combinations of the content that respondents ask their participants to verbalize.

Prompting participants: When using the classic think-aloud protocol [32], moderators are required to keep the interaction with their participants to a minimal level and only remind them to keep talking if they fall into silence. I asked respondents whether they prompt their participants during think-aloud sessions and found that only 22% of the respondents (n=37) keep the interaction minimal and do not prompt their participants with questions. In contrast, 78% of the respondents (n=132) prompt their participants.

In addition, 91% of the respondents (n=154) also reported how the frequency of prompting their participants had changed compared to when they just started their UX career and the result is shown in Figure 11. Among these respondents, 44% (n=67) felt that the frequency with which they prompt their participants remains *roughly the same*; 41% (n= 64) felt that the frequency for prompting their participants had only *slightly* changed; and 15% (n=23) felt that the frequency has changed *significantly*.

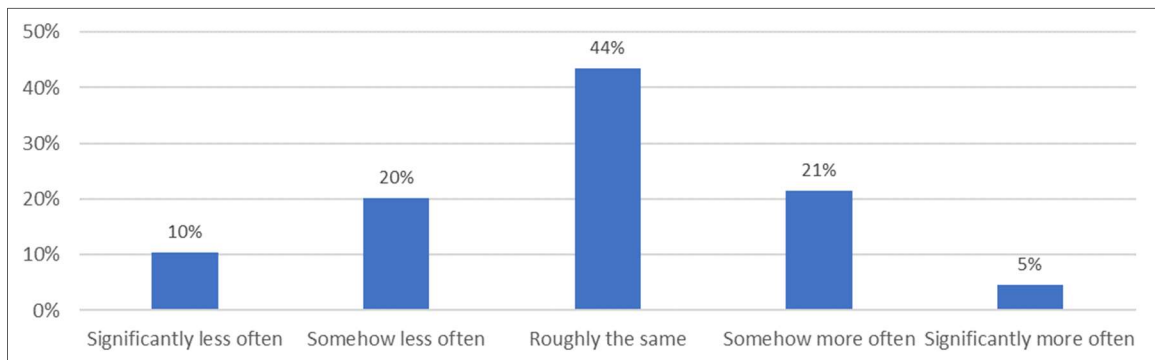


Figure 11. How the frequency with which respondents prompted their participants during think-aloud sessions had changed compared to when they just started their UX career.

Correlation analysis: I further examined whether there was any correlation between respondents' profile info and their practices of using think-aloud protocols. Specifically, we performed Spearman's rank-order correlation test when both variables were ordinal data and Chi-square test when there was a categorical data. Table 1 shows the results. In sum, the tests did not find any significant correlation for most pairs except between the size of respondents' companies and

whether respondents request their participants to verbalize content beyond what comes into the mind ($\chi^2(4, N = 169)=14.403, p=0.006$).

Table 1. Correlation analysis between responders' profile information and their practices of conducting think-aloud sessions (* indicates significance).

Respondents' profile info	Frequency of conducting practice sessions (ordinal data)	Whether asking users to verbalize content beyond what comes into the mind (categorical data)	Whether prompting users during the study session (categorical data)
The size of their companies (ordinal data)	$r_s(167)=-0.0294,$ $p = 0.7043$	$\chi^2(4, N = 169)=14.403,$ $p=0.006^*$	$\chi^2(4, N = 169)=1.3939,$ $p=0.8453$
The UX experience (ordinal data)	$r_s(167)=-0.0166,$ $p = 0.8308$	$\chi^2(4, N = 169)=2.6906,$ $p=0.6109$	$\chi^2(4, N = 169)=2.7057,$ $p=0.6082$

3.6.4 Analyzing think-aloud sessions

Activities performed for analyzing sessions: I asked respondents whether they performed the following activities when analyzing think-aloud sessions: review *observation notes of the usability test*; review the test session recording; review post-task interview data; review post-task questionnaire data; transcribe and review the transcript of the session. These options were based on a prior survey [64] and were updated via a pilot study. Figure 12 shows the result. Specifically, 89% of the respondents (n=151) review *observation notes*; 77% of them (n=130) review *the session recordings (e.g., audio/video recordings)*; 70% of them (n=118) review *post-task interview data*; 60% of them (n=102) review *the questionnaire/survey data (57%)*; and 56% of them *transcribe and review the transcripts (i.e., what participants said)*.

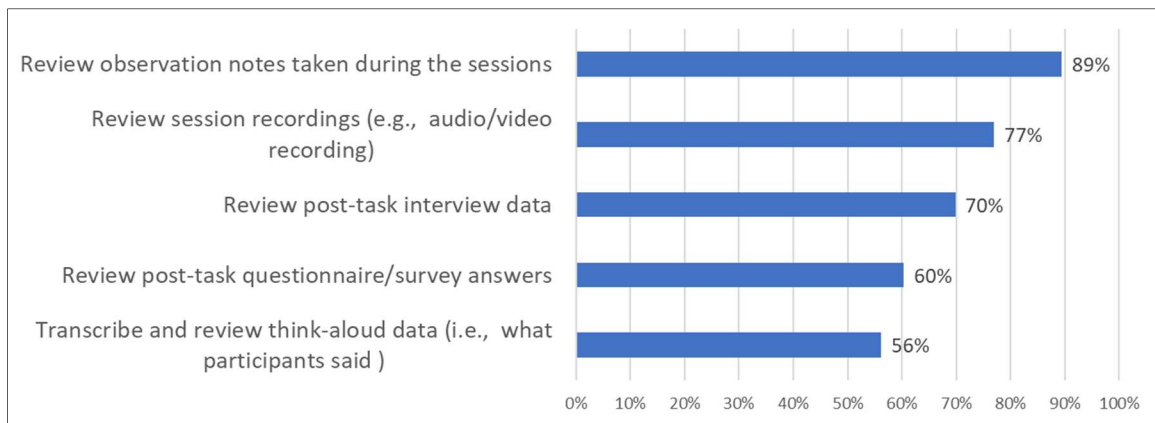


Figure 12. The activities that UX practitioners perform when analyzing think-aloud sessions.

Information for locating usability problems: Figure 13 shows the types of information that the respondents thought to help locate usability problems. Specifically, when reviewing think-aloud sessions to identify usability problems, 94% of them (n=159) thought *what participants were doing* (e.g., user actions on the interface) is helpful; 86% of them (n=145) thought *what participants said during the sessions* is helpful; and 76% of them (n=128) also thought *how participants said it* (e.g., pauses, tone) is also helpful.

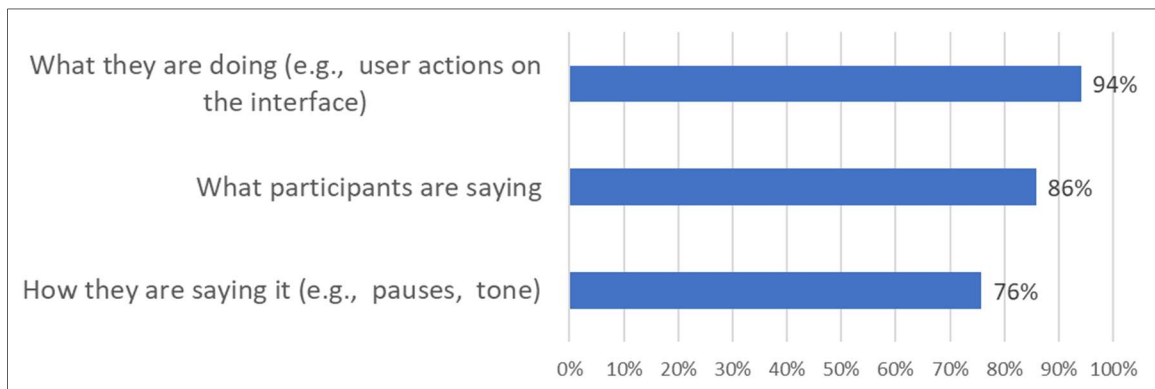


Figure 13. The types of information that help locate usability problems.

Information sought out from users' verbalizations: Figure 14 shows the information that the survey participants look for when analyzing their participants' verbalizations (i.e., utterances).

Specifically, 94% of them (n=153) looked for *expressions of feelings* (e.g., *excitement, frustration*); 89% (n=145) looked for their *participants' comments* (e.g., *feedback*); 74% (n=119) looked for their *participants' action descriptions*; 70% (n=116) looked for their *participants' explanations*; and 30% (n=49) looked for their *participants' design recommendations*.

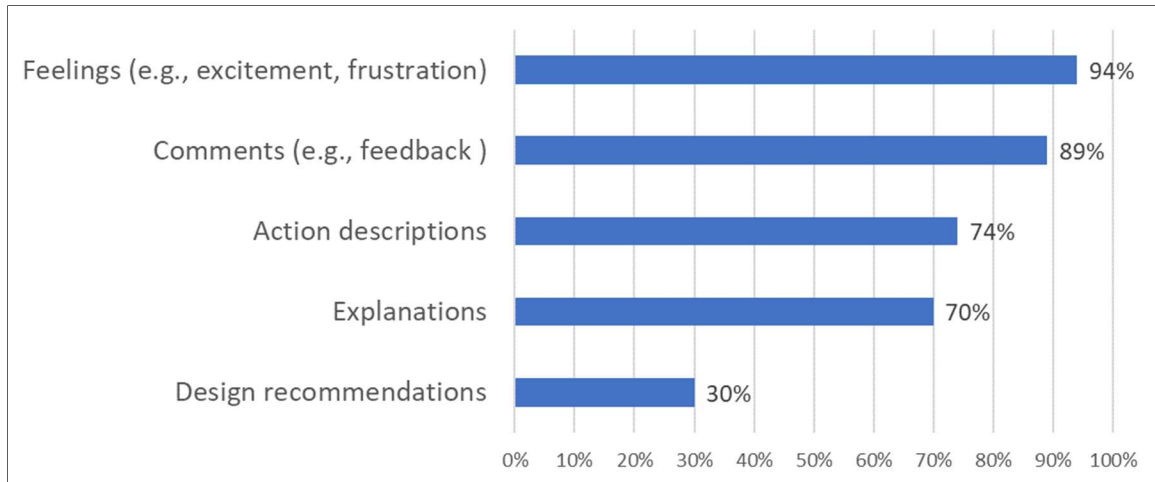


Figure 14. The types of information that UX practitioners seek in users' verbalizations.

Delivering analysis results: I asked respondents whether they performed the following three activities when delivering analysis results: *write an informal usability test report*; *write a formal usability test report*; *have a data analysis discussion meeting*. The survey did not provide definitions for these activities to make them open to interpretation. Respondents could choose multiple options if applicable. Figure 15 shows how the respondents deliver their analysis results. Specifically, when analyzing a think-aloud session, 69% of them (n=116) wrote an *informal usability test report*; 58% (n=98) wrote a *formal usability test report*; 57% of them (n=97) had a *data analysis discussion meeting*.

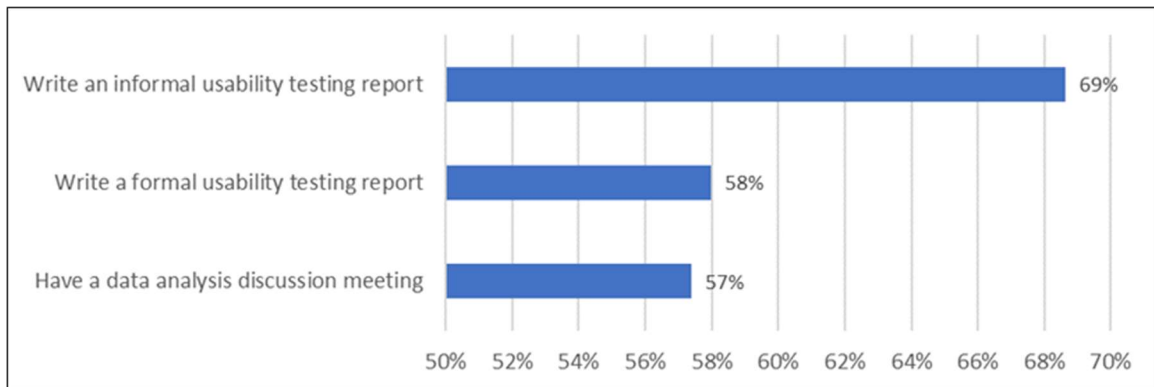


Figure 15. The ways in which UX practitioners deliver their analysis results.

Participation in the three types of data analysis: I asked respondents who would write formal and informal usability reports with the following six options: *only myself*; *UX designers/researchers*; *UX team lead*; *Lead of non-UX teams (e.g., engineering, marketing)*; *Other non-UX team members (e.g., engineers)*, and *C-level executives (e.g., CEO)*. In addition, I also asked respondents who would attend data analysis discussion meeting with the same set of options except “*by myself*.” They could choose multiple options if applicable. The result is shown in Figure 16. More than half of the respondents (56%, n=95) wrote informal usability testing reports alone and nearly half of the respondents (42%, n=71) also wrote formal usability testing reports alone. In addition, UX team members were the primary authors of informal/formal reports with occasional help from outside of the UX team. In contrast, non-UX team members were more involved in data analysis discussion meetings.

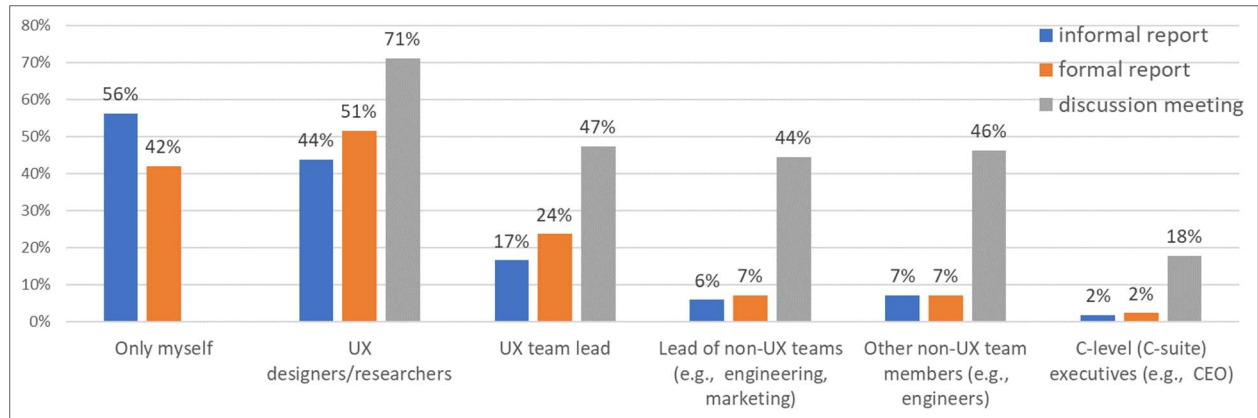


Figure 16. Participation in three types of data analysis activities: writing an informal usability test report; writing formal usability test report; having a data analysis discussion meeting.

3.6.5 Challenges with using think-aloud protocols

I asked respondents what their biggest inefficiencies or difficulties had been in conducting and analyzing think-aloud sessions an optional short-answer question and present the key findings from the responses below.

Challenges with conducting think-aloud sessions: The qualitative analysis reveals three main challenges related to conducting think-aloud sessions. First, getting their participants to think aloud is a challenge. Participants' personality and their ability to verbalize thoughts and the complexity and duration of the tasks are factors that influence the amount of content that they verbalize. For example, some people tend to be able to verbalize more readily than others, which can create an unbalanced representation of potential users. For some products, the target population may not be able to verbalize properly, e.g., children. Participants may also feel less comfortable verbalizing their thoughts when the task is complex. Furthermore, it may also be fatiguing for users to verbalize their thoughts if the task takes too long to complete.

Another challenge facing UX practitioners is to create a comfortable and neutral environment that encourages participants to verbalize their thought processes honestly. This is challenging because

participants might want to say nice things or may be reluctant to offer criticism during the test sessions, which could preclude UX practitioners from identifying usability bugs.

Finally, being patient and knowing when to interrupt participants is challenging. It is valuable to observe and understand how participants deal with the task themselves and recover from errors. Interrupting the process with prompts too early could change their way of interacting with the interface. Moreover, because part of the goal of usability evaluations is to gather data on what is difficult/impossible for users, it is often necessary to observe users struggle a bit during the evaluation to understand their “pain points”. However, it is also bad to let participants get stuck for too long as they may get frustrated, which may, in turn, affect the rest of the test session and consequently the amount of feedback that can be gained from the session.

Challenges with analyzing think-aloud sessions: While previous research reported general practices in analyzing usability evaluation [34], this survey study found specific challenges in analyzing think-aloud test sessions. The study showed that respondents review *session notes* (89%) more often than the *think-aloud session recordings* (77%) (see Figure 12). Respondents felt that reviewing think-aloud session recordings was arduous because recorded think-aloud sessions often contain so much data that transcribing and coding them takes a significant amount of time. Consequently, instead of transcribing sessions and reviewing transcripts, respondents often rely on “*their memory of participants’ sentiments and actions*” or the notes.

Despite the convenience of observation notes, respondents realized that it is “*easy to make judgments that might be off if they don’t refer back to actual transcripts or recordings*” and thus considered reviewing think-aloud session recordings a necessary part of their analysis process. First, because observation notes tend to be short, it is necessary to match the observation notes with the corresponding segments in the recordings to understand their contexts. Second, it is necessary to review the session recordings to capture points that might have missed by observation notes because notetakers can only write down the points that seem to be important from their perspective and any individual perspective can be incomplete or biased. Indeed, previous research

also suggested that while some of the usability problems may be captured by notes, much of the insight is often lost and needs to be reconstructed by conducting video data analysis later [53].

This survey study further identifies two challenges associated with reviewing think-aloud sessions. One challenge is to compare users' verbalization data with other streams of data to triangulate the issues that users encountered. One such comparison is to pair the user's actions on the interface with what they are saying (i.e., utterances) during the session. In scenarios where multiple streams of data are acquired, respondents had to correlate the verbalizations with other sensor data. Recent research has shown that considering verbalizations with other sensor data, such as eye-tracking data [29,30], EEG data [39] or functional Near-Infrared Spectroscopy (fNIRS) [61], can potentially increase the reliability and validity of the findings. Another challenge is to match the observation notes with the context in which the notes are taken. It is not always possible to note the exact timestamps when notes are taken. Consequently, matching notes (e.g., observations about users' facial expressions) with the audio stream of a recorded session often require evaluators to watch the entire recording. Another example of this challenge comes from the emerging VR and AR applications. To make sense of users' verbalizations when they interact with a VR or AR application, evaluators need to correlate the verbalizations with the visual content that participants observed during the sessions.

Reviewing think-aloud sessions is time-consuming. The survey respondents reported that they often had limited time to complete the analysis and faced the tension between achieving high reliability & validity in their analysis and completing their analysis efficiently. To cope with the tension, respondents reported using strategies, such as developing better note-taking skills or having a team of UX professionals observe a think-aloud test session and then socialize a recap session afterward.

In addition to reviewing recorded think-aloud sessions, respondents also pointed out that it can be valuable to keep track of the examples of different types of usability problems that they had observed over time and develop a taxonomy to describe the patterns in the data that commonly occur when users encountered usability problems. Such patterns, examples, and the taxonomy

could act as templates that potentially help them quickly identify common issues that users encounter and the solutions that they had accumulated in a new test context.

3.7 Discussion

The survey respondents worked in different geographic locations, in different industrial fields, and different sized UX teams. They also played different roles and possessed different levels of experience as UX professionals. Thus, the survey responses uncover a wide range of UX practitioners' practices surrounding the conduct and analysis of think-aloud sessions. Next, I discuss the implications of the survey responses.

3.7.1 General use of think-aloud protocols

Most of the participants (86%, 169 out of 197) use think-aloud protocols when conducting usability tests, which is viewed by the respondents as the most popular method to detect usability problems. Among the 91% of all respondents (n=179) who learned think-aloud protocols, 95% (169 out of 179) actually use the protocols in their usability tests. This result is consistent with the result of the survey study conducted in 2011 [64], which showed that 90% of the usability practitioners often use think-aloud protocols.

The study shows that concurrent think-aloud protocols are much more popular than the retrospective think-aloud protocols among UX practitioners. Approximately 91% of the respondents use the concurrent think-aloud protocols in at least half of the usability tests (see Figure 5). In contrast, only 37% of the respondents use the retrospective think-aloud protocols in at least half of the usability tests.

The study also reveals that think-aloud protocols are almost equally widely used in both controlled lab studies and remote usability testing. Compared to the most recent survey study conducted by McDonald et al. [64], this survey study identifies that remote usability testing is increasingly popular and think-aloud protocols are widely used in the remote usability testing. Research has shown that remote synchronous usability testing is virtually equivalent to the conventional lab-based controlled user studies [4]. In contrast, although remote asynchronous usability testing may

reveal fewer problems than conventional lab-based user studies, it requires significantly less time and thus is cost-effective [12]. Furthermore, although remote usability testing poses more workload on participants than the conventional lab-based user studies, participants generally enjoy the remote usability testing [14].

3.7.2 Conducting think-aloud sessions

To ensure the validity of participants' verbalizations, Ericsson and Simon provide three guidelines for conducting classic think-aloud sessions: keep the interaction to the minimal (i.e., only remind users to think aloud if they fall into silence for a period of time); use neutral instructions (i.e., instructions that do not ask for specific types of content); and have practice session(s) [32]. A meta-analysis of 94 think-aloud studies shows that an artificial change in performance can happen if these guidelines are breached [35]. However, previous research has documented that the gap between the theory and the practice of using think-aloud protocols existed [10] and the survey study provides evidence that such gap between the theory and the practice still exists. Specifically, we found that respondents did not always adhere to the three guidelines [32] and analyzed potential reasons for violating each guideline in the following paragraphs.

The study shows that only 16% of the respondents do not prompt their participants with questions except reminding them to keep talking when they fall into silence for a substantial period. Previous research has attributed the reason for not adhering the guidelines to the differences between the original goal of think-aloud protocols and the goal of using them in usability testing. The original goal is to study the unaltered human thought processes. Numerous studies have shown that probing or intervention (i.e., interaction with participants) could potentially alter the participant's thought processes, which could make the reported verbalizations not be an authentic representation of their thoughts [2,32,35]. Thus, UX practitioners should keep their intervention or probing to the minimum if possible. However, the goal of using think-aloud protocols in usability testing is mainly to identify usability bugs or to evaluate potential users' performance instead of just acquiring unaltered thought processes. Because of this difference, previous research suggests that UX practitioners may deviate from the guidelines and interact with the participants in two

situations [70]. One is when participants are frustratingly stuck. In this situation, interacting with them to help them recover from the error would allow the test to continue again, which would, in turn, allow UX practitioners to identify further usability issues. Another situation is when participants are struggling with a familiar problem, whose impact has been identified and well understood with previous test participants. In this situation, it is less meaningful to sit and observe participants be struggling with the problem again. Furthermore, as previous research suggested that audio interruptions and probing during think-aloud sessions may affect participants more than visual interruptions and probing [42], future research should examine the possibility of probing participants through the visual modality to acquire richer data while minimizing the risk of altering their thought processes.

Despite the guidelines recommending practitioners to use neutral instructions (only asking participants to report the content that naturally comes into their mind), the study reveals that only 7% of the respondents adhere to this guideline. Most of the respondents explicitly ask their participants to verbalize other types of content, such as feelings, comments, actions, and even design recommendations. This is concerning because research explicitly asking participants to verbalize a particular type of content can change their task-solving behavior [65], which may mask potential usability problems.

The study also shows that respondents also do not always follow the third guideline. For example, most of the participants almost never ask their participants to practice thinking aloud before conducting the actual sessions. Previous research shows that without practicing thinking aloud, participants often have difficulty verbalizing their thought processes [16]. Consequently, instead of treating the practice session as a burden, UX practitioners should treat it as an opportunity to help their participants to become familiar with thinking aloud, which in turns would help them verbalize their thoughts more naturally and frequently. This would ultimately help UX practitioners acquire more rich data to understand their participants' thought processes.

3.7.3 Analyzing think-aloud sessions

When analyzing think-aloud sessions, UX practitioners review observation notes more often than the session recordings and the transcriptions. One potential reason is that transcribing and reviewing the session recordings is arduous and time-consuming. Previous research pointed out that UX practitioners often face time pressure for their analysis [19]. Indeed, qualitative feedback provided by our survey respondents echoed this finding. Although the survey respondents largely knew that their judgments might be inaccurate if they did not refer to the actual session recordings, they often had to make trade-offs between achieving high reliability and validity and being efficient in their analysis. Currently, there are no known methods to deal with this tension effectively. The methods that survey respondents used include developing better note-taking skills and referring to the notes during analysis or having multiple UX practitioners observe a test session and socialize a recap session afterward. However, it remains unknown whether these methods are effective or if there are other more effective methods available. Indeed, recent research also suggested gaining a richer understanding of the tradeoffs that evaluators make and the impact of their decisions [62]. Therefore, future research should investigate methods and processes that can better balance the reliability, validity, and efficiency of the analysis of think-aloud sessions.

The study also reveals a need to identify common patterns from users' data that point to the moments when they experience problems in think-aloud sessions. As research has shown that users' verbalizations can be classified into different categories [24,40], it is worth exploring whether users tend to verbalize certain category (or categories) of content when they experience problems. Similarly, *do users tend to verbalize in certain ways (e.g., intonation, pitch, speech rate) when they experience problems?* Future research should explore whether such patterns exist. If these patterns do exist, they could be leveraged to design systems that automatically highlight portions of a think-aloud test session in which the user more likely experienced a problem, which in turn could help UX practitioners better allocate their attention during analysis.

3.8 Summary

I conducted an international survey study to understand the practices and challenges of using think-aloud protocols in the industry. Based on the responses from 197 UX practitioners who work in different industrial fields and different geographic locations, I have identified the practices and challenges surrounding the conduct and analysis of think-aloud sessions. The findings of the survey study could potentially inform UX practitioners about how their peers perceive and use think-aloud protocols. More importantly, this survey study also reveals opportunities in developing better methods and tools to make analyzing think-aloud sessions more effective. Specifically, it is valuable to explore *whether there are patterns in users' data (e.g., verbalizations, actions, and physiological measures) that commonly occur when they encounter problems* in think-aloud sessions (i.e., RQ2), *how to design computational methods that automatically detect portions of a think-aloud test session in which users were more likely to experience problems* (i.e., RQ3), and *how to leverage the automatically inferred usability problem encounters to facilitate UX practitioners' analysis process* (i.e., RQ4).

In the next three chapters (i.e., Chapter 4, Chapter 5, and Chapter 6), I will describe three studies that were designed to progressively identify and validate users' verbalization and speech patterns that tend to occur when they encounter problems. I will then describe the computational methods that leverage these verbalization and speech patterns to detect usability problems automatically in Chapter 7. Lastly, in Chapter 8, I will discuss the design, implementation, and evaluation of the intelligent visual analytics tool, VisTA, that presents automatically inferred usability problem encounters to UX practitioners to better understand how UX practitioners would perceive, interact, and integrate the ML-inferred usability problem encounters into their analysis of recorded think-aloud sessions.

Chapter 4 Study 1: Verbalization and Speech Features and Usability Problems

How are *users' verbalizations and speech patterns* linked to the *usability problems* that they experienced in concurrent think-aloud sessions? In other words, do participants tend to verbalize their thought processes with particular patterns when they encounter problems? I designed and conducted Study 1 to explore this question.

Study 1 examined if and how users' verbalizations and speech features could be used to identify usability problems. The study consisted of two phases: one to curate a dataset of think-aloud sessions and one to assess how verbalizations and speech features were used to analyze those sessions. In the first phase, I first conducted and recorded classic think-aloud sessions. In the second phase, I recruited usability evaluators to identify usability problems by reviewing these recorded think-aloud sessions. Each evaluator was provided with a tool for reviewing a session's audio recording, for visualizing verbalizations, and for logging usability problems. At the end of the study, I conducted semi-structured interviews to understand how evaluators made use of the tool and verbalizations to identify usability problems.

This chapter and the upcoming two chapters together answer the second research question of this dissertation:

RQ2: What are the subtle patterns in what participants verbalize (i.e., verbalizations) and how they verbalize it (i.e., speech features) that tend to occur when they encounter problems in think-aloud sessions?

4.1 Concurrent Think-Aloud Data Collection

4.1.1 Participants

I recruited four participants (3 females, aged 19-24) from a student social group at a local university to participate in think-aloud sessions. To reduce any language issues that might interfere with their verbalization process, all participants were native English speakers.

4.1.2 Procedure

I conducted the think-aloud sessions by following Ericsson and Simon's three guidelines to encourage participants to make valid verbalizations [32] (see Section 2.2). The three guidelines are as follows: arranging a practice session for participants to practice and get used to thinking aloud; using neutral instructions to ask participants to verbalize all the thoughts that come into their mind without requesting any specific type of content; and keep the intervention to the minimal by only reminding participants to keep talking if they fall into silence for a long period of time (e.g., 15 seconds). First, the moderator described the study details to the participant and played a short online video tutorial [75] on the classic think-aloud protocol. Afterwards, each participant was asked to perform three think-aloud sessions, using the primary functions of three devices: to set an alarm clock to *ring one hour from now*, program a coffee machine (De'Longhi BCO264B) to *prepare two cup of strong-flavored drip coffee for 7:30 in the morning*, and to *copy two single-sided sheets of paper to a double-sided sheet of paper using a photocopier* (Brother DCP-L2520DW). The alarm clock was given as a *practice* trial. The coffee machine and photocopier were chosen particularly because they were representative of devices that people may use on a regular and occasional basis, respectively. No participants had experience using these specific models.

Prior to conducting think-aloud sessions, the moderator explicitly asked each participant to "say out loud everything that you say to yourself silently. Just as if you are alone in the room speaking to yourself" [32,65]. During the study, the moderator sat quietly away from the participant's view, to monitor the study. The moderator remained silent and did not interact with participants except to remind them to keep talking if they remained silent for longer than 15 seconds.

To ensure that there was an equal number of participants using each device first, I randomized the order of the two devices given. For each think-aloud session, the participant was given the task instructions, the device, and its printed instruction manual. The moderator also explained the tasks to the participants and answered any questions to ensure that they understood the study procedure.

The study lasted about an hour, and each participant was compensated with \$20. All of the eight think-aloud sessions were audio-recorded and lasted between 516 and 769 seconds ($M=683$, $SD=129$).

4.2 Analysis of Think-Aloud Sessions

4.2.1 Participants (Evaluators)

I recruited 12 participants (8 females) to analyze the eight think-aloud sessions that were collected in the concurrent think-aloud data collection phase of the study. The participants were 22-31 ($M=25$, $SD=2$) years old. All participants were screened to have basic knowledge of usability testing methods and hands-on experience in using think-aloud methods (seven were graduate students in a UX design master's program, one was a UX staff in a startup, and the rest four had taken HCI/UX courses). There were no foreseen issues with the use of student evaluators for the study, as previous research has shown that students, who have been taught and practiced the protocol in school, are well-suited for identifying problems in think-aloud sessions [71]. Henceforth, these study participants are referred to as *evaluators* to distinguish them from the *participants* who took part in the concurrent think-aloud data collection phase.

4.2.2 Study Design

Each evaluator analyzed two think-aloud sessions (of the eight), assigned at random. The study was designed to ensure that 1) each evaluator analyzed two think-aloud sessions performed by different individuals using two different devices; 2) half of the evaluators analyzed a coffee machine-related think-aloud session first. With this assignment mechanism, each session was analyzed by three evaluators.

I counterbalanced the two sessions assigned to each evaluator so that they were about two subjects (ID: 1-4) using two devices (coffee machine: c, printer: p). For example, 1-c denoted the session for which subject 1 used the coffee machine. In addition, the 8 sessions assigned to 4 of the evaluators were as follows: 1-c, 4-p; 3-p, 2-c; 4-c, 1-p; 2-p, 3-c. The counter-balancing was similar for the rest of the evaluators.

4.2.3 Verbalization Categorization and Speech Features Extraction

Verbalization Categorization: The first task for each evaluator involved listening to audio and categorizing the verbalizations into the four categories. I built a tool to facilitate this process (Figure 17). Inspired by a popular online transcribing tool, oTranscribe [112], I designed the following functions in the categorization tool: press ESC key to play or pause the audio; press F1/F2 button to rewind/fast-forward the audio. I also provided the ability to speed the audio up 1.5x/2x/2.5x. The tool was designed to minimize the number of operations that evaluators were required to carry out. Whenever the audio is played/paused, the start/end timestamp of the audio segment is automatically recorded. Using the radio selector on the tool's interface, evaluators only needed to select an appropriate category for each audio segment. I followed a similar categorization approach used in previous work [24,30] by asking evaluators to divide the audio into segments and assign each audio segment with one of the four categories. Audio segment borders were determined by pauses between verbalizations and the content of these verbalizations, following the same procedure used in the literature [24,30]. Each audio segment corresponded to a verbalization unit, which could include single words, but also clauses, phrases, and sentences. However, it was ultimately up to evaluators as to how they created their segments. Figure 17 shows the tool's interface. The definitions of the four categories were always displayed on the interface. The colored bar at the bottom showed the category labels that are already assigned by the evaluator.

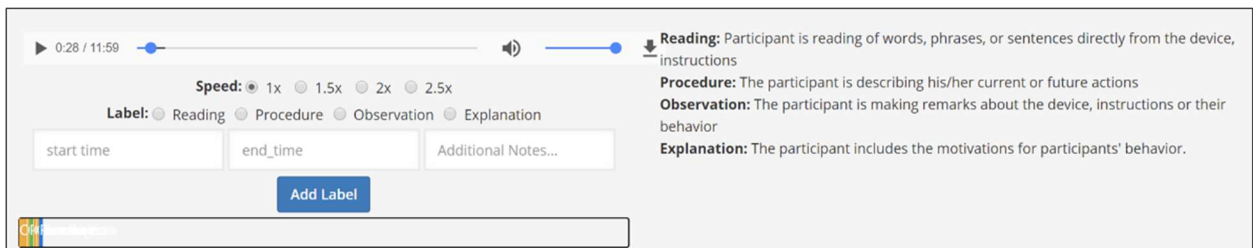


Figure 17. The tool for evaluators to segment audio recordings of think-aloud sessions and provide category labels for each segment. The colored bar at the bottom shows the category labels that are already assigned by the evaluator.

Speech features: I used the Web Speech API [90] to automatically generate transcripts of think-aloud session recordings. To address the shortcomings in automatic speech recognition, the errors in automatic speech recognition were corrected manually. The tool also provides the think-aloud audio recordings and their transcripts as inputs to the Penn Phonetics Lab Forced Aligner (P2FA) [108]. P2FA computes the start and end time of each transcribed word, which is used to highlight the transcribed word corresponding to the cursor's position when the cursor is on any of the voice feature visualization panels (the timestamp corresponding to the cursor's position is extracted and compared with each word's start and end time). The audio-aligned transcript also provides the start and end time of pauses between each word. These pauses were silent periods. The start and end time of the verbal fillers were manually labeled. I used a state-of-the-art sentiment analysis model, VADER [48], to compute the sentiment of each sentence in a transcript.

To compute the speech rate, for every N ($N=30$) aligned entities (including words and short pauses between words) in an audio-align transcript, the speech rate is then computed as the number of words N divided by the duration of these N aligned entities. A 50% overlap between each two adjacent N aligned entities was used to smooth the computed speech rate. As the automatic speech recognition algorithm [90] removes verbal fillers, the start and end times of these fillers were manually labeled by the authors. The analysis tool also uses a speech processing toolkit, Praat [9], to compute loudness (dB) and pitch (Hz) over time.

4.2.4 Tool for analyzing Think-Aloud Sessions

To facilitate evaluators to identify usability problems, I designed and implemented visualization and analysis functions in the prototype tool (Figure 18) that visualizes the category information that usability evaluators and the six speech features on aligned and synchronized timelines. These features were designed to guide usability evaluators to visually navigate the audio, observe patterns, compare features, and log usability problems. The tool reads the *category* labels and the *speech features* from files and visualizes them in seven aligned timelines (Figure 18). Figure 19 shows the call-out views of the tool's interface.

The cursor is shown and synced in all seven panels, and the corresponding word at that timestamp is highlighted in the transcript. Clicking on any point of a visualization panel brings the audio to the corresponding timestamp, and subsequent pressing of the ESC key plays the audio from that timestamp. Dragging the cursor on any feature panel highlights a portion of the visualization. The background color of the selected portion in all features panels turns grey to indicate the highlight. After highlighting, pressing ESC plays the audio from the start of the highlighted portion. Because longer silences may reveal different information about the verbalization than shorter ones, the tool also provides five length filters (> 1s, 3s, 5s, 10s and 15s) to allow usability evaluators to selectively focus on longer- or shorter- duration silences (Figure 20).



Figure 18. The tool for evaluators to analyze a recorded think-aloud session. It visualizes the transcript of the session on the left panel. Seven audio features are represented as charts on the right panel. Highlighting any part of a chart will highlight the corresponding transcript on the left panel. The bottom part of the tool allows an evaluator to log problems and select the verbalization and speech features that they used to identify the problems.



Figure 19. Callout views of three parts of the analysis tool: seven feature panels (top); highlighted word of the current timestamp (middle); problem description area (bottom).

The bottom of the tool provides functions for logging usability problems. To ease the logging of the start and end time of a usability problem for evaluators, the tool automatically detects and fills these two timestamps whenever usability evaluators highlight a portion of any chart or the transcript. Inspired by previous work [58,67], I used a structured problem report that included a description of the user problem, the problem’s context and verbalization features that indicated the problem. Specifically, the text fields and checkboxes on the bottom part of the UI allows evaluators to describe usability problems and select the verbalization features that they used to identify them. To better visualize the temporal relationship between usability problems and all visualized features, a colored segment will be visualized on the “Problem” timeline when a problem is added.



Figure 20. The visualization of the silence (the colored part of the timeline) before and after selecting a silence length filter.

4.2.5 Procedure

To understand how information about usability problems can be inferred from the audio recordings of think-aloud sessions alone, evaluators were not given access to the devices and instruction manuals. The study facilitator told the evaluator that they were to review audio-recorded think-aloud sessions to identify problems that participants encountered. The facilitator informed each evaluator that there were two steps to evaluating each of the two think-aloud audio recordings. The *first* step was to use a tool (Figure 17) to divide the audio into small segments and label each segment with one of the four categories: reading, procedure, observation and explanation. The categories were based on the literature [24] and adjusted slightly to better fit the tasks in the study: Reading: *read words, phrases, or sentences directly from the device or instructions*; Procedure: *describe his/her current/future actions*; Observation: *make remarks about the device, instructions or themselves*; Explanation: *explain motivations for their behavior*. The *second* step involved using

a tool (Figure 18) to identify usability problems. The definition of a usability problem that was adopted was “anything that interfered with a user’s ability to efficiently and effectively complete tasks” [52] and evaluators were asked to consider any aspect of the products that might cause confusion, frustrations and hamper the user’s ability to use them. The details of the two steps are explained in the next two subsections. Figure 21 shows that an evaluator was analyzing a recorded think-aloud in the study.

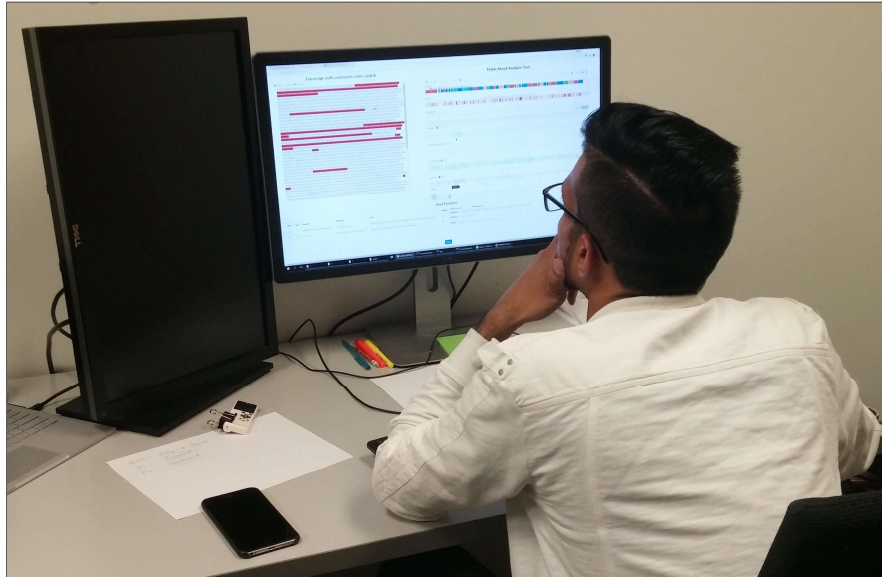


Figure 21. An evaluator was analyzing a recorded think-aloud session during the study.

After each evaluator completed the two steps for the two think-aloud audios that they were given, I conducted semi-structured interviews to understand how they used categories information and speech features. Each evaluator had 30-40 minutes to complete the task related to each assigned audio. The entire study ran for about 1.5 hours. Each participant was compensated with \$30.

4.3 Analysis and Results

In addition to the verbalization category labels and usability problems from 12 evaluators, the study also collected qualitative feedback from evaluators about how they identified problems and used verbalization and speech features in the semi-structured interviews.

4.3.1 Problems Identified by Usability Evaluators

Two UX researchers independently validated usability problems that were identified by evaluators by checking their problem descriptions and listening to the audio segments corresponding to the time period of the problems. Disagreements were resolved via discussion. The result shows that 170 of the logged problems were valid. The average number of problems identified per evaluator per device was as follows: the coffee machine ($M=6$, $SD=2$) and copier ($M=8$, $SD=4$). A paired-samples t-test was conducted to compare the number of problems identified for each device. There was no significant difference in the number of usability problems identified by the evaluators for each device ($t(11) = -1.23$, $p = .24$).

4.3.2 Verbalization Categories and Identified Problems

Two UX researchers followed the same strategies as used in previous studies [24,30,46] to count the number of verbalized thought units in each category and compute each category's percentage. As each audio segment corresponded to a thought unit, two researchers counted the number of audio segments that were labeled with a category for all categorization files. Results show that Observation (O) (35%), Reading (R) (32%) and Procedure (P) (25%) were frequently used but Explanation (E) (8%) appeared much less frequently.

Next, the number of times that each verbalization category was associated with the identified usability problems was computed. Specifically, for each usability problem logged by a usability evaluator, its start and end time were identified. The researchers went through the categorization labeled by this evaluator and counted the number of times that each category appeared in this time interval. The researchers repeated this process for all usability evaluators' data. Figure 22 shows an example of the usability problems and the verbalization categorization analyzed by a usability evaluator. For example, the second usability problem (pink color) was associated with three verbalization categories (Procedure, Reading, and Observation) once each.

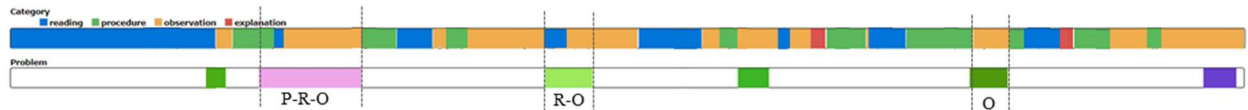


Figure 22. The verbalization categories for a think-aloud recording and the problems identified by an evaluator.

The percentage of each verbalization category appearing in the segments that were associated with usability problems: Observation (31%), Reading (31%), Procedure (25%), Explanation (13%). The researchers performed a repeated-measures ANOVA on the number of times that each verbalization category was associated with usability problems (the sphericity assumption was not violated). The results show that there was a significant difference ($F(3, 33)=14.53, p =.00, \eta_p^2=.57$). Post hoc tests revealed that: 1) there was a significant difference among the following three pairs: Explanation and Observation ($p=.004$), Explanation and Reading ($p=.001$), Explanation and Procedure ($p=.01$); 2) there was no significant difference between any of the following three categories: Reading, Observation, and Procedure.

The researchers further analyzed the number of categories associated with each usability problem. The results show that 37% of the usability problems were associated with a single category (Observation: 39%, Reading: 30%, Procedure: 30%, Explanation: 1%) and the Observation category was the most likely associated with usability problems among all.

4.3.3 Speech Features and Identified Problems

The researchers counted the number of times that each feature was used by usability evaluators to identify usability problems that were logged by the analysis tool and the result is shown in Figure 23. In addition to the features that are related to the actual content verbalized (i.e., transcript, sentiment, and verbalization category), how participants verbalized (e.g., the use of verbal fillers, silence, speech rate, pitch, and loudness) were also used by evaluators to identify usability problems. The result of the Mauchly's test shows that the sphericity assumption was violated ($\chi^2(27) = 54.97, p = .002$), therefore the degrees of freedom was corrected. A repeated-

measures ANOVA with a Greenhouse-Geisser correction found no significant difference in the number of times that each feature was used ($F(3.26, 35.88) = 2.12, p = .11, \eta_p^2 = .16$).

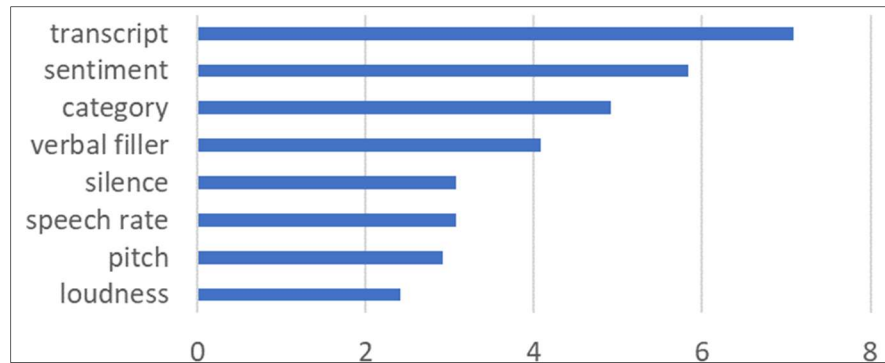


Figure 23. The number of times that each feature that evaluators used to identify problems.

4.3.4 Qualitative Feedback

I present the results of the semi-structured interviews, which aimed to gain a deeper, more detailed understanding of how evaluators used verbalization features to identify problems.

Categories. In line with the quantitative data, half of the evaluators mentioned that the audio segments labeled Observation was the most useful category for locating usability problems. They explained that Observation segments often contained users' opinions and comments about the devices, instructions and their experience that helped evaluators understand the user's thoughts or struggles. Evaluator 8 (s1_ev8) noted, "[*Observation*] is usually when users express their feelings when they are doing something. It shows if they had a doubt about the task or were frustrated".

Although the *Reading* category may intuitively seem to be passive as users are just reading instructions, the quantitative data shows that it is sometimes also indicative of usability problems. Usability evaluators often considered long periods of consecutive reading to be associated with usability problems: "reading a lot is usually a good indicator of them having problems understand[ing] the instructions."-s1_ev3; "If you see a significant amount of reading consecutively, that's bad. Because that means the person had to constantly revisit the instructions. Why should users have to constantly revisit the instructions if the design of the product and

*instructions were good?”-s1_ev7; “when they pause while reading, that probably means they are hesitating.”-s1_ev9. One evaluator considered *Procedure* to be the most useful category, as such segments helped identify what users were doing. This helps assess the fluency of the user’s interactions. For example, “if you see a lot of *Procedures*, it usually means users are just going no problem at all.”-s1_ev3. *The Explanation* was the least used category. This was partially because it was similar to the *Observation* category and hence few segments were assigned as this category. “*Explanation is a bit hard to find because I had to search for keywords, such as ‘because’ to locate them. Sometimes they did not use ‘because,’ they just explained it right away, so I had to go back one or two sentences to identify it.*”-s1_ev10*

Almost half of the evaluators (5/12) explicitly stated that patterns of categories, especially those that showed *repetitive* attempts, were the most useful for identifying usability problems. S1_ev6 explained, “if the user was reading, then commented something, and then went back to reading the same thing, it might indicate confusion or a memorization problem.”

Evaluators consistently felt that *segmenting and labeling the audio recordings was time-consuming and overwhelming*. Rather than dividing thought units in fine granularity, they would produce much longer segments to save time.

Silence. Usability evaluators noted that periods of silence were useful in identifying user confusion. However, they also felt that the interpretation of this feature was highly context dependent. For example, silence can occur when users were operating on a machine, thinking, quietly reading and comprehending instructions. As it is not useful having only the knowledge of when a user is silent, evaluators often navigated to times slightly before and after a silent period to assess its context and make a more informed judgement.

Verbal Fillers. Unlike silence, evaluators felt that verbal fillers were more indicative of problems. Two main strategies for interpreting verbal fillers were adopted. The first involved identifying big blocks of verbal fillers in the search for problems. Another way to use it in combination with the silence feature. In particular, audio segments that included the “verbal fillers-silence-verbal fillers”

pattern was strongly indicative that the user had encountered a problem. However, evaluators noted that verbal fillers alone do not necessarily suggest usability problems as verbal fillers usage is common in speech and is subject to individual differences.

Sentiment. During the interviews, evaluators expressed that the sentiment of an audio segment was most useful for predicting usability problems. Most evaluators said that they focused on large blocks of negative sentiment or large variances in the sentiment graph.

Evaluators also used negative sentiment in conjunction with verbal fillers and silence and mentioned that it was a strong indicator that the user was having trouble in accomplishing their goals. However, evaluators also mentioned that sentiment information was not always accurate. Based on evaluators' feedback, the current algorithm used to determine sentiment is limited in detecting sarcasm, and as a result, the sentiment of some audio segments could be mislabeled.

Speech Rate. Because the speech rate may vary from individual to individual, like verbal fillers, it is hard to be certain whether or not a user's speech rate may indicate a usability problem. However, some evaluators still noted that those segments, where the speech rate slows down dramatically, could indicate that users were thinking, confused, or interpreting instructions. In contrast, other usability evaluators found that a high speech rate could indicate problems as well. For example, one evaluator mentioned that one user tended to read instructions very fast when she was experiencing difficulties, such as when she attempted to locate the photocopier's "copy" button. This user also repeatedly regurgitated instructions to ensure that she read everything correctly. Big clinches in the speech rate line graph, which shows large variations in a short period of time, were also indicative of problems (Figure 24 left).

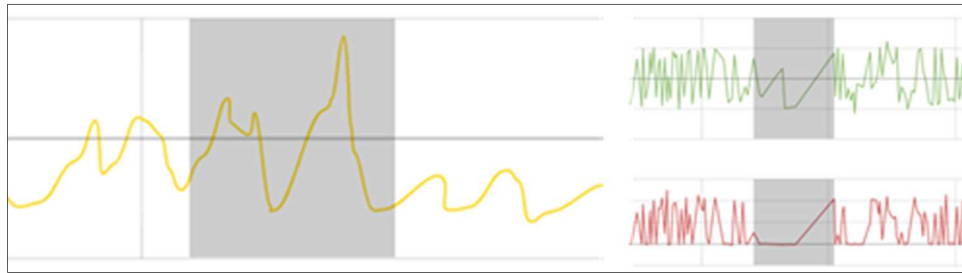


Figure 24. (left) clinches on the speech rate graph; (right) irregular patterns appeared in both loudness and pitch graphs.

Loudness & Pitch. Usability evaluators felt that it was harder to find useful information from the graphs for loudness and pitch than other features because these two graphs were considered to be much noisier, consistent with the fact that these two features were less frequently used in logs (Figure 23). One evaluator explained as follows, “*because the lines are so busy and there are so many ups and downs, it is hard to pick up significant areas.*”-s1_ev7. Some evaluators felt that the loudness was not as trustworthy because variances in loudness might be attributed to noise in the surrounding environment, rather than the user’s speech. However, they also noted that loudness and pitch were in sync and the erratic patterns in loudness and pitch graphs could indicate periods of confusion. Figure 24 right shows such an example, in which the lines for loudness and pitch were sparser compared to other regions of the graphs.

Transcript. While verbalization categories were able to show high-level information about what type of information was verbalized and speech features were able to point out patterns in how users verbalized, the transcripts were useful in that they revealed the detailed contents of verbalization. It was mainly used by usability evaluators to track the user’s progress (see the highlighted text in Figure 18). Furthermore, it was also used to obtain direct quotes from users as evidence of usability problems.

4.4 Summary

In this study, I examined the relationship between *verbalization categories* and *speech features* and *usability problems*. Results show that audio segments that are labeled as the *Observation* category and that are of *negative sentiment* are more indicative of usability problems. Evaluators often used the category and the speech features (i.e., categories, sentiment, fillers, silence, speech rate, pitch, and loudness) in conjunction with each other to identify usability problems.

After conducting the study and reflecting on it, I identified three limitations in the study design. First, I only recruited *a small number of participants* for curating the concurrent think-aloud dataset—more participants and usability evaluators could potentially increase the validity of the results.

Second, I asked the evaluators to first listen to the recorded think-aloud sessions and segment and categorize the content into four categories before analyzing the sessions. However, given the limited time available, evaluators found it challenging to segment and label the recorded think-aloud sessions thoroughly. To save time for the subsequent analysis (i.e., identifying usability problems), evaluators tended to identify long segments, and thus included parts of the audio that should have been excluded. For example, because it was hard to isolate the brief comments right after users read instructions or performed procedures, they often grouped such comments together, which should be labeled as *Observation*, into surrounding context as either *Reading* or *Procedure*. Thus, there would likely be more *Observation* segments than reported in Study 1 *if the categorization process was done without time constraints*.

Third, evaluators found that presenting the audio transcripts as a *continuous block of text* (Figure 18) was not visually-friendly, and some suggested to add line-breaks after the text of each transcribed thought unit. To address the three limitations and further validate the findings of Study 1, I designed and conducted a second study (Study 2), which will be introduced in the next chapter.

Chapter 5 Study 2 (Confirmation Study)

I designed and conducted Study 2— a confirmation study— to validate the findings of Study 1 by addressing its limitations. Specifically, I made the following changes to the study design, the analysis tool, and the analysis method based on the study observations and the feedback from the UX evaluators in Study 1:

- I recruited more participants to participate in the think-aloud sessions to curate a larger think-aloud dataset; I also recruited more usability evaluators to analyze these think-aloud sessions.
- I reduced the workload of the usability evaluators by releasing them from annotating the verbalization category information so that they could better focus on identifying problems. Instead, the verbalization category was annotated by researchers prior to the study.
- I revised the analysis tool to incorporate UX evaluators' feedback from Study1 to make it more user-friendly for evaluators.
- I analyzed the links between verbalization and speech features and the encounters of usability problems by using precision, recall, and F-measure, which take the proportion of each feature into consideration and thus result in more objective measures than the simple percentage or counting measures.

5.1 Concurrent Think-Aloud Data Collection

5.1.1 Participants

I recruited a new set of eight participants (five females and three males, aged 19-24) from a student social group at the university to participate in the think-aloud sessions. All the participants were native English speakers to avoid language issues that might interfere with the verbalization process. Participants were from various graduate and undergraduate programs, including design, life science, cell biology, cognitive science, computer science, occupational therapy, psychology

and cinema studies. This diverse background was chosen to reduce the biases inherent to any particular discipline.

5.1.2 Procedure

This study followed the same procedure as Study 1 with the exception that different testing devices were used: a coffee machine (*De'Longhi BCO264B*), which is used by people regularly, and a universal remote control (*RCA RCRN03BR*), which is used relatively infrequently. A new set of devices were purposely chosen to assess the generalizability of the findings. The same alarm clock and the task as in Study 1 were used in practice trial. All participants had not used these specific models prior to the study. The ordering of the two devices given participants was counter-balanced to ensure that there was an equal number of participants using each device first. For each think-aloud session, participants were given the device, the hard-copy of its instruction manual, and a task to perform. Table 1 shows the tasks, which involved using each device's primary functions. For the universal remote task, participants were given a DVD player and a TV to carry out the task. Each participant was compensated with \$20 for the hour-long study.

Table 2. Tasks that participants worked on the two products.

Device	Tasks
Coffee machine	program the coffee machine to make two cups of strong flavor drip coffee at 7:30 in the morning.
Universal Remote	program the universal remote control to operate a DVD player.

All think-aloud sessions were audio recorded. As each participant performed two think-aloud sessions using different devices, there were a total of 16 think-aloud sessions. The average duration of the sessions was 891 seconds ($SD=222$) for the coffee machine and 649 seconds ($SD=100$) for the universal remote control. The average duration of all sessions was 770 seconds ($SD=208$).

5.2 Analysis of Think-Aloud Sessions

5.2.1 Participants (Evaluators)

I recruited 16 participants (12 females) to analyze the think-aloud audio recordings that were collected. Ages ranged from 20-28 ($M=24$, $SD=2$). Apart from two participants who worked in

the industry as UX evaluators, all participants were either graduate students in UX programs or senior year undergraduate students who had previously taken UX courses in the university. All participants had previously conducted and analyzed think-aloud sessions and reported the number of projects, as part of a course, internship or job, for which they had employed the think-aloud method is as follows: 1-5 projects (12), 6-10 projects (2), and > 10 projects (2). These participants are again referred to as *evaluators* to distinguish them from those who participated in the concurrent think-aloud data collection.

5.2.2 Study Design

I counter-balanced the think-aloud sessions assigned to each evaluator so that 1) each evaluator analyzed two think-aloud sessions for distinct devices and users; 2) half of the evaluators began the study with a coffee machine think-aloud session. With this study design, each think-aloud session was analyzed by exactly two evaluators. Table 3 shows the counter-balancing scheme that was employed in the study.

Table 3. The counter-balancing scheme (p1-p8 denote the participants' IDs in the think-aloud data collection).

Evaluator ID	1st session	2nd session	Evaluator ID	1st session	2nd session
	Coffee machine	Universal Remote		Universal Remote	Coffee machine
1	p1	p2	9	p1	p5
2	p3	p4	10	p3	p7
3	p5	p6	11	p2	p6
4	p7	p8	12	p4	p8
5	p2	p3	13	p5	p2
6	p4	p5	14	p7	p4
7	p6	p7	15	p6	p1
8	p8	p1	16	p8	p3

5.2.3 Verbalization Categorization and Speech Features Extraction

In Study 1 (Chapter 3), I experimented with automatic speech recognition (i.e., Web Speech API [90]) to generate transcripts of think-aloud sessions and found that it lacked accuracy and

substantial effort had been devoted to correct automatic transcription errors. Thus, in this study (i.e., Study 2), all think-aloud recordings were manually transcribed to ensure their accuracy.

Two coders followed a similar approach used in previous work [24,30] to divide each audio recording into small audio segments and assign each audio segment with one of the four **verbalization categories**: *reading, procedure, observation, and explanation*. The definitions of the four verbalization categories were the same as the previous study, which were based on the literature [24] and adjusted slightly to better fit the tasks in the study: Reading (R): *read words, phrases, or sentences directly from the device or instructions*; Procedure (P): *describe his/her current/future actions*; Observation (O): *make remarks about the device, instructions or themselves*; Explanation (E): *explain motivations for their behavior*. The beginning and end of an audio segment were determined by pauses between verbalizations and the content of these verbalizations, following the same procedure used in the literature [24,30]. Each audio segment corresponded to a verbalization unit, which could include single words, but also clauses, phrases, and sentences.

The level of agreement between the two coders was assessed by computing the inter-rater reliability (IRR) for a single think-aloud session. The IRR score came out to be sufficiently high (Cohen's kappa: $k=0.91$). For the audio segments that the coders labeled differently, they discussed and resolved disagreements. The remaining audio recordings were then labelled separately by the two coders.

Next, I computed six voice features from each think-aloud audio recording: **sentiment, speech rate, loudness, pitch, silence, and verbal filler**. The sentiment of each audio segment was computed using VADER [48], a state-of-the-art sentiment analysis model. The speech rate for each audio segment was computed by dividing the number of words spoken in an audio segment by the segment's duration. Loudness (dB) and pitch (Hz) was computed using the speech processing toolkit Praat [9]. The start and end times of each period of silence and verbal filler were manually labeled to ensure the accuracy. These voice features and the verbalization category labels

were loaded and displayed in a tool that was designed to assist usability evaluators in identifying and logging usability problems.

5.2.4 Tool for Analyzing Think-Aloud Sessions

The tool was designed based on the previous version used in Study 1 (Figure 18) and improved based on the feedback from the evaluators in Study 1. Figure 25 shows the interface of the updated tool. The left panel visualizes the transcript of an audio recording, with each line corresponding to one audio segment. The rest of the interface was the same as the one shown in Figure 18.

5.2.5 Procedure

Prior to the start of the study, the facilitator first introduced the tool's function, how to use it, and then gave each evaluator a few minutes to familiarize themselves with the tool. The facilitator informed evaluators that they would use a tool to identify problems that users were experiencing in the audio recordings. Each evaluator had a maximum of 30 minutes to analyze each of the two think-aloud audio recordings that were assigned to them. After analyzing the audio recordings, I conducted semi-structured interviews to understand how evaluators identified problems and made use of verbalization features. The entire study lasted for about 1.5 hours. Each evaluator was compensated with \$20.



Figure 25. The updated tool for evaluators to analyze recorded think-aloud sessions. It visualizes the transcript of the recording on the left panel (a), one line per audio segment. The seven features (i.e., category, silence, verbal fillers, sentiment, speech rate, loudness, pitch) are represented as time-synchronized charts on the right panel (b). Selecting any part of a chart highlights the corresponding transcript on the left panel. The bottom of the tool (c) allows for describing usability problems and the features used to identify the problems.

5.3 Analysis and Results

5.3.1 Number of Labels per Verbalization Category

For all the 16 think-aloud audio recordings, the number of times that each verbalization category was used by the evaluators were counted. Table 3 displays this information for each device separately and in tandem. Notably, the four categories were used in similar proportions for both devices.

Table 4. The frequency and percentage of the audio segments labeled with each verbalization category.

Device	Verbalization category			
	Reading	Procedure	Observation	Explanation
Coffee machine	276 (29%)	235 (25%)	371 (40%)	52 (6%)
Universal remote	185 (28%)	177 (27%)	256 (39%)	33 (5%)
All devices	461 (29%)	412 (26%)	627 (40%)	85 (5%)

5.3.2 Problems Identified by Usability Evaluators

In total, 273 problems were identified by the usability evaluators, 148 in the think-aloud sessions for the coffee machine and 125 for the universal remote control. Two researchers validated each problem that was logged, by checking the problem description and listening to the corresponding audio segment independently. Disagreements were resolved via discussion. Of these problems, seven were assessed to be invalid because they missed proper description, or their descriptions did not match with the content of the associated audio segment. With these problems removed, a total of 266 problems were considered in the subsequent analysis.

The average number of problems identified per evaluator for each device was: coffee machine ($M=9$, $SD=3$) and universal remote ($M=8$, $SD=3$). A paired-samples t-test was conducted to compare the number of problems identified for each device. There was no significant difference in the number of problems identified between the two devices ($t(15) = 1.54$, $p = .15$).

5.3.3 Verbalization Categories and Identified Problems

The problems that were logged by the usability evaluators were analyzed to understand how the four verbalization categories are related to the problems. First, for each problem logged by an evaluator, the number of different verbalization categories that fell into the problem's start and end times was counted. Figure 22 shows the audio recording of a think-aloud session with labeled verbalization categories (top) and the problems identified by an evaluator (bottom). For example, the first problem (pink color) was associated with the three categories (i.e., *Procedure*, *Reading*, and *Observation*), which occurred once each.

To better understand the correlation between verbalization categories and usability problems, I computed the *precision* and *recall* of each verbalization category in locating usability problems using the following equations:

$$precision = \frac{\text{the number of segments labelled as a particular category associated with an identified problem}}{\text{the total number of segments labelled as the same category in the entire session}}$$

$$recall = \frac{\text{the number of segments labelled as a particular category associated with an identified problem}}{\text{the total number of segments associated with an identified problem}}$$

I used precision and recall as measures because they *account for the base rate of each verbalization category* in a think-aloud session when considering their relationship with usability problems.

Precision and recall can be used to answer the following two questions:

- If an evaluator randomly checks a segment labeled with a particular category, what is the chance of finding a problem?
- If an evaluator checks all segments labeled with a particular category, what percentage of problems could be found?

The greater precision of a verbalization category indicates that evaluators would have a higher chance of finding a problem by examining a segment labeled as the category and the greater recall of a verbalization category indicates that evaluators would be able to catch more problems if they examine segments labelled as the category. Furthermore, to assess the overall relevance of a verbalization category with usability problems, I further computed the F-measure, which combines precision and recall as a single measure using the following equation: $\frac{2*precision*recall}{precision+r}$. Figure 26 shows the precision, recall, and F-measure of each verbalization category in identifying usability problems. It shows that while the segments labeled as *Observation* are the most relevant to usability problems, the segments labeled as *Explanation* are the least relevant to usability problems. The segments labeled as *Procedure* or *Reading* are also relevant to usability problems, but less so than the ones labeled as *Observation* and more so than the ones labeled as *Explanation*.

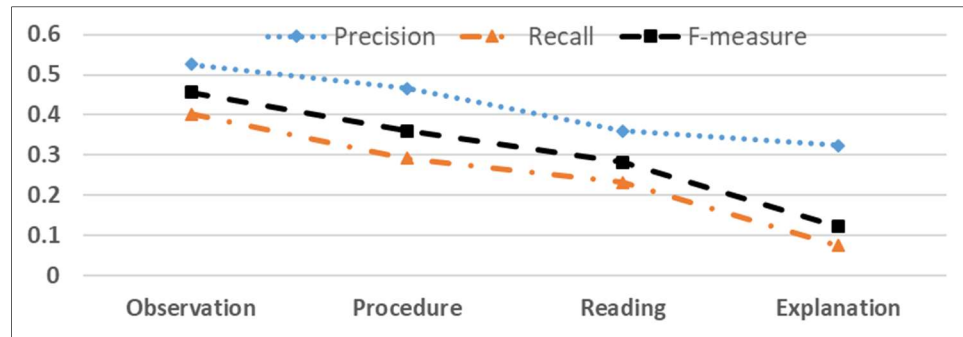


Figure 26. Precision, recall, and F-measure of each verbalization category in identifying problems.

To examine if the trend shown in Figure 26 still holds for different devices respectively, I computed the precision, recall, and F-measure of the four categories in identifying problems for each device separately. Table 5 shows these measures for each device respectively. The numbers in each column under each one of the three measures in Table 5 generally decrease, which indicate that the same trend that was observed in Figure 26 largely holds for each device separately.

Table 5. Precision, recall, and F-measure of each category in identifying problems for each test device.

Category	Precision		Recall		F-measure	
	Coffee machine	Universal remote	Coffee machine	Universal remote	Coffee machine	Universal remote
Observation	0.54	0.52	0.40	0.40	0.46	0.45
Procedure	0.47	0.46	0.30	0.28	0.37	0.35
Reading	0.37	0.35	0.23	0.24	0.28	0.28
Explanation	0.29	0.37	0.07	0.08	0.11	0.13

To understand how evaluators used the combination of verbalization categories in identifying problems, I further computed the precision, recall, and F-measure of twelve pairs of verbalization categories in identifying problems. For example, the pair “R-O” refers to one *Reading* category segment or an uninterrupted sequence of the *Reading* category segments followed by one *Observation* category segment or an uninterrupted sequence of the *Observation* category

segments. The category pairs are mutually exclusive. Table 6 shows the results, which suggest that the verbalization pairs that are most relevant to usability problems typically contain the *Observation* category and the verbalization pairs that are least relevant to usability problems contain the *Explanation* category. Comparing Table 5 and Table 6, it is evident that the *Observation* category was more relevant to usability problems than any verbalization pairs.

Table 6. Precision, recall, and F-measure of the verbalization category pairs in identifying problems.

Verbalization category pair	Precision	Recall	F-measure
R-O / O-R	0.27	0.46	0.34
P-O / O-P	0.26	0.33	0.29
P-R / R-P	0.21	0.11	0.15
O-E / E-O	0.30	0.05	0.08
P-E / E-P	0.18	0.04	0.06
R-E / E-R	0.16	0.01	0.02

5.3.4 Speech Features and Identified Usability Problems

Usability evaluators also selected the features that they used in helping them find usability problems by checking the checkboxes of the corresponding features in the area C of the tool (Figure 25). Figure 27 shows the number of times that each feature was used by all evaluators. All the verbalization and speech features were used by evaluators to identify usability problems. Among them, *category* and *sentiment* were the most frequently used features, while *pitch* and *loudness* were the least. The result of the Mauchly's test shows that the sphericity assumption was violated ($\chi^2(27) = 56.57, p = .001$), therefore the degrees of freedom was corrected. A repeated-measures ANOVA with a Greenhouse-Geisser correction found a significant difference ($F(3.43, 51.37) = 6.18, p = .001, \eta_p^2 = .29$). The results of the post-hoc pairwise comparisons show no significant difference between pairs except the following: category and loudness ($p = .012$), verbal filler and transcript ($p = .004$), silence and transcript ($p = .005$), pitch and transcript ($p = .000$), loudness and transcript ($p = .003$).

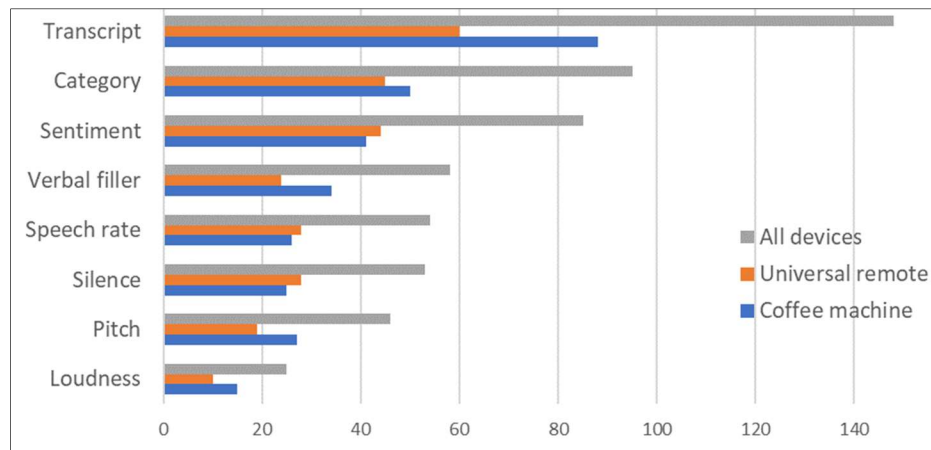


Figure 27. The number of times that each verbalization and speech feature was used by evaluators for finding usability problems.

5.3.5 Qualitative Feedback on the Use of Verbalization & Speech Features

Two researchers transcribed the interviews and coded the transcripts independently. They further discussed to consolidate their codes. The interviews provide a deeper, more detailed understanding of how evaluators used *verbalization* and *speech features* to identify problems.

Category. Evaluators underscored that the Observation category was most indicative of problems (“*Observation describes how the users were feeling and how they commented their confusions*”-s2_ev6). Some evaluators relied on segments labeled Observation to help them focus on finding problems quicker (“*I know that most of the problems aren’t going to be in Reading or Procedure. Instead, they would be in Observation.*”-s2_ev15). Moreover, some found that Observation audio segments with a long duration signaled a problem (“*When users are confused, they do a lot more Observations and sometimes explanations. You’ll see less of the Reading and Procedure.*”-s2_ev7). Evaluators also noted that Observation category contained a diverse amount of information, which is not necessarily related to problems (“*It could be users expressing a problem but could also be them commenting something worked*”-s2_ev1). This feedback is consistent with the quantitative measures in section 5.3.3.

Evaluators generally thought that the Reading category was tied less to problems, mainly because *“reading was just users reading instructions”-ev13*. On the other hand, some evaluators noted that the Reading category was still useful in indicating problems. For example, a user who is confused about a set of instructions may repeatedly regurgitate them (*“I noticed that if they were confused, they tended to read the instructions more. If they knew what they would do, they would intuitively work through it.”-s2_ev1*; *“Repetitions mean that they try to say or do the same thing over and over again in a short period of time. When someone is experiencing difficulties, you might see such repetitions.”-s2_ev7*).

While the Observation category was the most useful category for identifying problems, evaluators emphasized that category combinations helped provide context as to why users were encountering problems (*“When the problem presents itself, it is usually in the Observation category. But the real problem was usually already there for a while. You need to go back and read or listen to the segment before the Observation to understand the context.”-s2_ev13*; *“Reading becomes important to understand a comment when the user expressed an Observation after Reading, such as ‘oh, I don’t understand...’”-s2_ev10*).

Sentiment. When using the sentiment information, evaluators mostly focused on audio segments with negative sentiment (*“I mostly checked the low part of the sentiment chart. When they are unsatisfied or confused, they naturally tend to say negative words, which would be the low part of the chart.”-s2_ev15*). Evaluators gauged user sentiments by examining the transcript as well (*“Sentences with negative sentiments, such as “it sucked”, were the ones that I tried to find while reading the transcript. The sentiment is important but having to look at the transcript and sentiment chart at two different places is a bit distracting.”-s2_ev2*).

Apart from using a negative sentiment as a place to look for problems, evaluators also paid attention to sudden changes in sentiment (*“‘I feel like they should do that unless...Oh, No, OK.’ When there are two words back and forth that one is negative and the other one is neutral or positive. It means that they changed their tone immediately, which usually indicate their confusions”-s2_ev14*). In essence, abrupt transitions in sentiment might be Eureka moments (or

Aha! moments) for the user [5], i.e., the sense of suddenly coming to an understanding of a concept that was previously confusing. In usability testing, Eureka moments might imply that a product does not follow users' intuitions and is likely not easy to use.

Evaluators also noted that a shortcoming of the sentiment voice feature is that it is based solely on the contents of a verbalization (i.e., what was said) and does not give insight into how such verbalizations are made (*"It is possible that the same content can mean different things until I listen to it."*-s2_ev1). As a result, the sentiment feature was sometimes inaccurate because it failed to consider one's tone of voice, which may at times be more telling of a user's emotions rather than what was verbalized, as in the case when users are sarcastic. For example, sentences like *"oh, that's helpful"* may be negative in actuality, but be classified as positive using the text-based sentiment analysis in the study. Thus, evaluators suggested that listening to audios can be important to assess the true sentiment of a sentence.

Verbal Fillers. Some evaluators mentioned that users would use more verbal fillers right before and during the presence of problems. Rather than using the verbal filler chart, evaluators reported that they primarily used the transcript to look for the verbal fillers. Many evaluators had expressed a desire for them to be made more visually salient in the transcript, such as by highlighting them. Evaluators also noted that they could not rely only on verbal fillers in making judgments about usability problems, as people's use of them can vary widely: some people may use verbal fillers sparingly, while some people may use them habitually. To gain a sense of users' manner of speech, evaluators suggested engaging in a conversation with them prior to a think-aloud session.

Words such as *"what?"*, *"where?"* and *"how?"* were also considered to be verbal fillers and evaluators found them to be useful in identifying audio segments with problems (*"There are certain things that you can say to show your confusion without literally say 'I'm confused.' For example, you may use 'huh' or ask questions, like what? Where? These words mean that you are confused. Otherwise, you wouldn't be asking questions."*-s2_ev14).

Speech Rate. Like verbal fillers, speech rate varies from individual to individual and is therefore difficult use as a telltale sign of usability issues. In spite of this, some evaluators noted that a lower-than-normal speech rate may indicate that users were thinking, interpreting instructions or were confused (*“I looked at parts of the chart that were below the average, because when the user in the first session I analyzed had a problem, she spoke slower.”-s2_ev5*). Moreover, a higher-than-normal speech rate could also indicate problems. One evaluator mentioned that the user of one think-aloud session was reading the instructions very fast when she had trouble finding the right content. However, high speech rate can be unreliable, since users may speak quickly even though they are not encountering problems.

Silence. Evaluators also made use of periods of silence in verbalizations as a sign of usability issues. In particular, evaluators took advantage of the filtering function (Figure 1) to look for prolonged periods of silence (>3s) that may suggest user confusion (*“I felt that with 1 second filter, there are too many left [on the chart]. With 3 seconds, there are a reasonable number of silences for me to analyze.”-s2_ev9*).

Similar to speech rate and verbal fillers, relying primarily on silence could result in making false conclusions. Users may fall silent for reasons other than usability issues, such as when they are operating a machine, thinking, or when quietly reading or comprehending instructions. To gain contextual information, evaluators reported examining audio segments occurring just before and after silent periods.

Pitch & Loudness. Pitch and loudness were the least used features. Evaluators felt that it was difficult to detect patterns in the pitch and loudness charts since they did not have much meaningful variation (*“The chart was mostly the same kind of looking, so it’s hard to tell exactly what’s meaningful.”-s2_ev5*). They, however, still believe that pitch and loudness can be useful to assess a user’s level of confidence or the state of confusion, such as decreasing their volume or raising their pitch (*“When users are losing confidence in what they are doing, the loudness of their voice tends to be lower.”-ev6*; *“Whenever a user ends a sentence with a higher pitch like asking questions, it has always been that he is confused.”-s2_ev14*).

Transcript. Evaluators reported that having access to audio transcripts saved time because it “got more into users’ head”-ev2 and allowed them to attend to important or interesting verbalizations without having to listen to the audio recording all the time. For example, they noted that they could easily skip irrelevant audio segments (“I skipped [listening to] the parts that I knew were just them describing what they were doing.”-s2_ev1) and focus on problematic segments (“I highlighted the part in the transcript that seems to be a problem and then listened to the audio and analyzed the charts on the right.”-s2_ev5). This feedback is consistent with the log data, which showed that on average, evaluators only listened to 70% of the think-aloud audios. Evaluators also expressed that the transcript helped identify verbal fillers and other remarks made by participants that were signs of usability problems, such as “I’m going to start this over again, or I’m stuck”-s2_ev9.

5.4 Summary

I present and discuss the findings of how verbalization categories and speech features relate to usability problems in this subsection.

Verbalization Categories. As evidenced by this study, usability evaluators benefitted from having access to an audio recording’s verbalization categories. Firstly, the results revealed that audio segments with the *Observation* category were more indicative of usability problems than other categories, presumably because these audio segments often described a user’s concerns about a product or their behavior. The proportion of segments with the *Observation* category was higher than that in Study 1, which was expected. Because the researchers who coded the categories before the study did not face time pressure in the categorization process as the evaluators did in Study 1, they were able to segment the recordings with more granularity. This had helped to isolate the brief comments right after users read or performed some actions. Additionally, the *Reading* category helped to pinpoint places where users had difficulties in making sense of instructions, as they would spend long periods of time reading instructions; often repeating the same set of instructions over and over again. Segments categorized as *Procedure*, as in prior work [66], helped evaluators understand and assess the ease at which users could follow a set of instructions.

Verbalization categories were also helpful in finding contextual information to understand the problems faced by users, particularly segments categorized with *Reading* or *Procedure*, as these segments described actions that users attempted to perform.

The segments that were least associated with problems contained the *Explanation* category. One reason could be that the audio segments with this category were low in general (5%). This number is in line with those reported in previous studies (*e.g.*, 5% in Cooke's study [24] and 7% in Elling *et al.*'s study [30]), perhaps implying that users tend not to explain or provide motivation for their behavior. One example of an *Explanation* category segment following a *Procedure* category segment from a universal remote control session was as follows: "*let's try the Auto Code Search [method] because it says it's the easiest method.*"

Notably, the pairs that were most closely associated with problems were the combinations of *Observation* (O) with either *Reading* (R) or *Procedure* (P). In particular, the accumulated recall of the top four pairs (*R-O*, *O-R*, *P-O*, *O-P*) was 0.79, which suggests that evaluators could find 79% of the problems when examining these pairs. This is perhaps because the context information provided by *Reading* or *Procedure* segments is needed to understand problems in *Observation* segments. In the ideal case where no problems are encountered, a user's verbalizations should alternate between *Reading* and *Procedure*. I posit that such pairs, in which users deviate from reading and performing procedures to make an observation, indicate that they may be facing difficulties. In addition, the likelihood that users are facing difficulty increases with the amount of deviation from reading and performing procedures to make an observation. However, further investigation is needed to confirm this speculation.

As shown in this study, the *Observation* category was the greatest telltale sign of problems, with around half of all audio segments containing the *Observation* category label being tied to a usability problem (see the precision values shown in Table 5). This result implies that with a roughly 50% rate of accuracy, usability evaluators can identify problems when randomly examining a segment labeled as the *Observation* category. As the recall values for the *Observation* category were also around .5, usability evaluators would find around half of the usability problems

if they only focused on segments labeled as the *Observation* category. The implication is that although the *Observation* category is the greatest telltale sign of problems, usability evaluators should also leverage other information to increase the chance of identifying usability problems. For example, for greater reliability when examining *Observation* segments, many evaluators suggested combining *Observation* and *negative sentiment* information, on the grounds that if an *Observation* segment is about something working as expected, the corresponding sentiment would not be negative. However, as text-based sentiment analysis is inaccurate, this approach still requires evaluators to refer to the corresponding audio segments.

Speech features. Evaluators found that all the voice features were useful, especially sentiment. They often used *sentiment* together with *category* (e.g., the *Observation* category and *negative sentiment*) to quickly focus on interesting segments of the transcript or audio. Regarding the visual design of the tool, evaluators expressed a desire for sentiment and verbal filler information to be combined with the transcript, as opposed to being visualized in separate charts, as integrating these features may reduce the spread of their attention on the tool's user interface. Evaluators also proposed other useful parts of speech that may indicate problems, such as when users ask questions (i.e., What? Where? How? Huh?). *Repetitive patterns*, such as reading a set of instructions over and over again or performing actions repeatedly, also raised red flags.

Because verbal fillers, speech rate, and silence tend to vary from individual to individual, evaluators felt that they would need to speak to the participants to get a sense of their normal speech patterns to use these features. The implication is that although these voice features are potentially useful to identify usability problems, knowing a user's colloquial speech habits (i.e., the baseline of the voice features) might help evaluators better leverage these features.

Chapter 6 Study 3 (Generalization Study)

I identified and validated the links between users' verbalization and speech features and the encounters of usability problems in think-aloud sessions through Study 1 and Study 2. In addition, I further identified *three factors* that may concern the generalizability of the findings of these two studies.

1) *Physical Devices vs. Digital Systems*. In both Study 1 and Study 2, physical devices were used for think-aloud sessions. Digital systems, such as websites, are another type of products that require extensive usability testing and have been used as test products for think-aloud related research (e.g., [2,24,109,110,30,40,41,55,65,66,81,95]). People operate digital systems (e.g., websites) differently than physical devices. Physical devices have fixed interfaces with a limited number of controls with which the user can interact. The challenge with completing tasks on physical devices might be figuring out how to map steps and actions to features and controls on the physical devices. In contrast, digital systems do not have those same physical constraints. The challenge here might be finding specific interface features to satisfy the user's need. Additionally, limb motion is often required for operating physical devices, while digital systems require more eye motion and relatively small-scale hand movement (e.g., operating a mouse).

2) *Verbalizations with Audio Recording vs. Verbalizations with Video Recording*. Study 2 validated and enriched the findings of Study 1 and provided a better understanding of how verbalization and speech features were used to identify usability problems. In these two studies, evaluators only assessed think-aloud sessions from their audio recordings to avoid the potential influence of other modality and to better assess the role of verbalization and speech features. Although evaluators were able to identify problems to a proficient degree from just the audio recordings, it would be interesting to explore whether including other modalities, such as video, might have added benefit or change the verbalization and speech patterns that indicate problems.

3) *Visualization of Verbalization and Speech Features*. In Study 1 and Study 2, usability evaluators had access to the visualizations of verbalization and speech features (e.g., verbalization

categories and speech features). This might have influenced their think-aloud analyses since visualizations might have directed their attention to certain parts of the sessions more often than others and subsequently led them to identify more or fewer problems.

To better understand whether these three factors affect the verbalization and speech features that are indicative of usability problems, I designed and conducted Study 3 (i.e., the generalization study). Specifically, I aimed to explore the following research questions:

Are verbalization and speech patterns that signal usability problems different for the *physical devices* and *digital systems*?

Are verbalization and speech patterns that signal usability problems different when the *video* recording of a think-aloud session is also provided?

Are verbalization and speech patterns that signal usability problems different when the *visualization* of verbalizations is not provided?

6.1 Concurrent Think-Aloud Data Collection

6.1.1 Participants

To curate a new think-aloud dataset, I recruited a new set of eight participants (four females and four males, aged 19-26), all of whom were native English speakers, from student social groups at a local university. Like Study 1 and Study 2, native English speakers were chosen to reduce language barriers. Participants had diverse backgrounds: biology, creative writing, environmental science, neuroscience, and pharmacology.

6.1.2 Procedure

The study's procedure was the same as the first two studies. The products tested in the think-aloud data collection included two websites in addition to the two physical devices, the coffee machine (*De'Longhi BCO264B*) and the universal remote control (*RCA RCRN03BR*): a national science and technology museum (*STM*) and a national history museum (*HM*) website. These two websites

were chosen as the participants could potentially be their users, and these websites possessed a certain number of usability problems, as was determined by a heuristic evaluation that I conducted. Three tasks that covered some of the target websites' main functions were identified and used in the think-aloud sessions. Table 7 shows the three tasks for each of the two websites.

Table 7 Tasks for the two websites used in think-aloud data collection.

Websites	Tasks
STM	Your friend is an 8th-grade science teacher. She asks you to check if there are any available school programs in April at the Science museum. Your task is to find out whether there are any programs that may be suitable for 8th-grade students in April.
STM	Your uncle has an 11-year-old child. One day, the child asks you a question, "what is it like to be a scientist or an engineer?" You've heard that the museum offers interactive presentations during which children can interact with speakers, who are scientists. Thus, your task is to find out if there is any such program in March for an 11-year-old child.
STM	You are a college student and are working on an assignment about early telescopes. Your task is to obtain a photo of an instruction manual, which is for an early telescope.
HM	Your friend is a 7th-grade teacher. She is organizing a trip for 30 7th grade students to the history museum. Your task is to help your friend find an available program in March for 30 7th grade students.
HM	Your friend has a 4-year-old child and is planning to take him to the history museum. Please help your friend to find out the number of activities that are appropriate for a 4-year-old child in March.
HM	You are a graduate student and are currently researching the topic of first peoples in Canada. Your task is to search for an essay on the topic.

All think-aloud sessions were both video- and audio- recorded. A 27" 4K monitor connecting to a laptop was placed on a desk. Participants performed all website tasks on the monitor. All website task sessions were screen captured with a picture-in-picture window of a participant's face using a Logitech HD Pro Webcam and the Open Broadcaster Software. The think-aloud sessions of participating using two physical devices (the coffee machine and the universal remote) were captured with two wall-mounted cameras, which monitored each participant's face and hand movements. For better quality audio, I used to a clip-on voice recorder instead of the camera's embedded microphones and later manually synchronized audio and video streams. Each participant was compensated with \$20 for the hour-long study.

In total, 64 think-aloud sessions were recorded (each participant performed eight think-aloud sessions: one task for each physical device and three tasks for each website). All sessions ranged

from 62 seconds to 1255 seconds ($M=360$, $SD=279$). The average duration of the sessions for each device or website was as follows: coffee machine ($M=854$, $SD=251$), universal remote control ($M=619$, $SD=195$), STM ($M=222$, $SD=131$), and HM ($M=247$, $SD=153$).

6.2 Analysis of Think-Aloud Sessions

6.2.1 Participants (Evaluators)

I advertised the study in several local UX/HCI social groups via Facebook and Slack. In total, 16 participants (11 females) were recruited to analyze the think-aloud sessions. These participants were referred to as *evaluators* henceforth, similar to Study 1 and Study 2, to distinguish them from the *participants* who took part in the concurrent think-aloud data collection. The evaluators' ages ranged from 22 to 50 ($M=27$, $SD=7$). Their self-reported professions were as follows: usability specialist (2), UX designer (2), UX researcher (1), graduate students specialized in UX (11). All of the participants had experience in the think-aloud method from their jobs, internships and graduate course projects. The number of projects for which they had used think-aloud method to conduct usability tests was as follows: 1-5 (3), 6-10 (11), and > 10 (2).

6.2.2 Study Design

I counter-balanced three factors—test products (i.e., the physical devices and websites), the modality of think-aloud recordings, and visualization—through a balanced Latin-square design so that each evaluator analyzed all four of the test products. Evaluators analyzed two of their sessions with the audio recording. For the other two, evaluators were given the video-recording, which came with the audio as well. Additionally, each evaluator only had access to the visualization in two of their four sessions. An example of the assignment mechanism for four usability evaluators is shown in Table 7.

Table 8 A balanced Latin Square design for every four evaluators.

Coffee machine	Remote	History Museum	Science & Technology Museum
Audio + Visualization	Video + Visualization	Video	Audio
Video + Visualization	Audio	Audio + Visualization	Video
Audio	Video	Video + Visualization	Audio + Visualization
Video	Audio + Visualization	Audio	Video + Visualization

To reduce potential carry-over effect between test products, I changed their order according to a 4×4 balanced Latin-square design for the rest evaluators (Table 8). The sessions assigned to each evaluator were also conducted by different think-aloud participants to avoid potential biases that might occur if they were to analyze the same think-aloud participant’s sessions more than once. Note that for each website, an evaluator analyzed three recordings, each one corresponding to a task in the Table 7.

6.2.3 Verbalization Categorization and Speech Features Extraction

Verbalization categories and the speech features (i.e., silence, verbal fillers, sentiment, speech rate, loudness, and pitch) for each think-aloud recording were generated following the same process described in Study 2. All these features were loaded and displayed on the updated analysis tool as described in the next section.

6.2.4 Tool for Analyzing Think-aloud sessions in Different Study Conditions

Based on the feedback from Study 2 and the purpose of this study, I updated the think-aloud analysis tool that was used in Study 2 (Figure 25) to show a different user interface for each experimental condition (Table 8): *Audio*, *Video*, *Audio + Visualization*, and *Video + Visualization*.

Figure 28 shows the interface when evaluators *only* had access to the *audio* recording of a think-aloud session, which included an audio player and controls to play and pause audio. Figure 29 shows the interface when evaluators *only* had access to the *video* recording of a think-aloud session, which included a video player and controls to play and pause the video. Figure 30 shows the interface when evaluators had access to both the *audio* recording of a think-aloud session and

the *visualization* of verbalization and speech features. The visualization of the verbalizations was the same as that of Study 2, which include a transcript, a verbalization category chart, and seven voice feature (i.e., silence, verbal fillers, sentiment, speech rate, loudness, and pitch) charts. Lastly, Figure 31 shows the interface when evaluators had access to both the *video* recording of a think-session and the *visualization* of the verbalizations. The left two columns are the same as the Audio + Visualization condition. The right column shows the video recording of a think-aloud session.

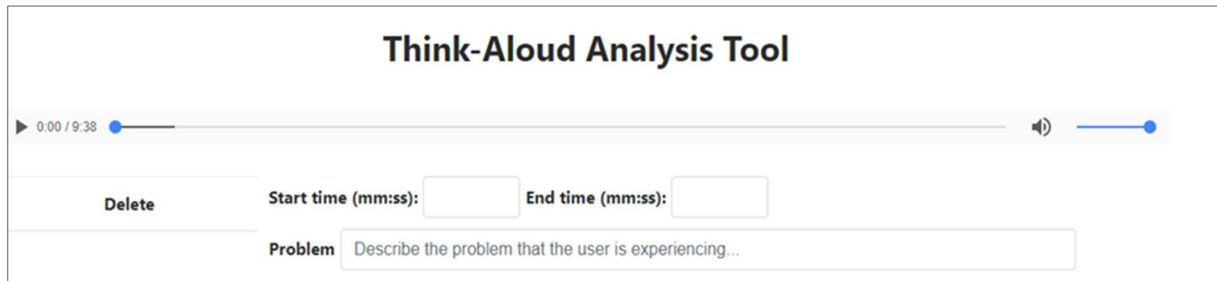


Figure 28. The think-aloud analysis tool's interfaces for the *Audio only* condition

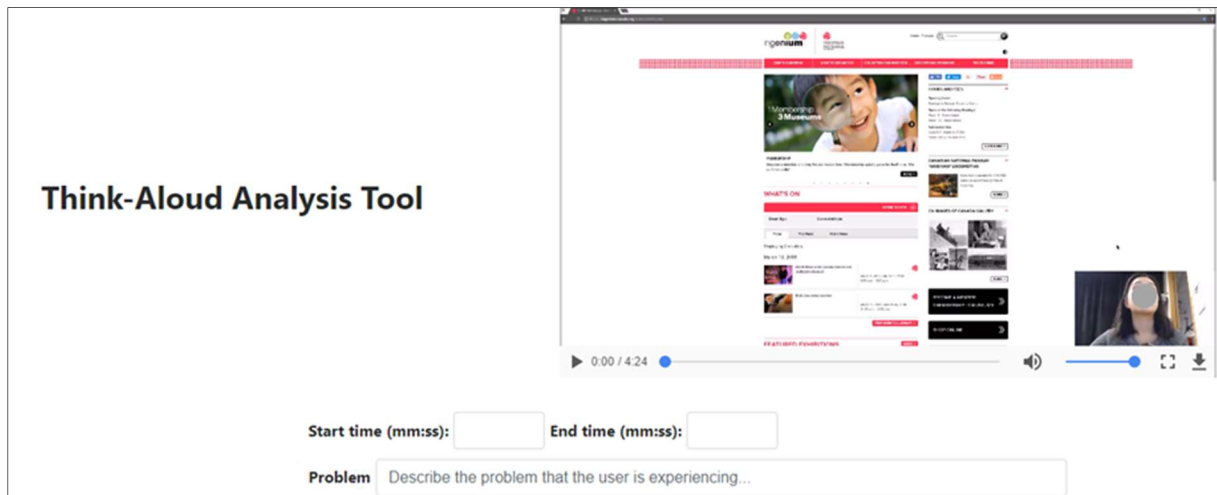


Figure 29. The think-aloud analysis tool's interface for the *Video only* condition.



Figure 30. The think-aloud analysis tool’s interface for the *Audio + Visualization* condition.

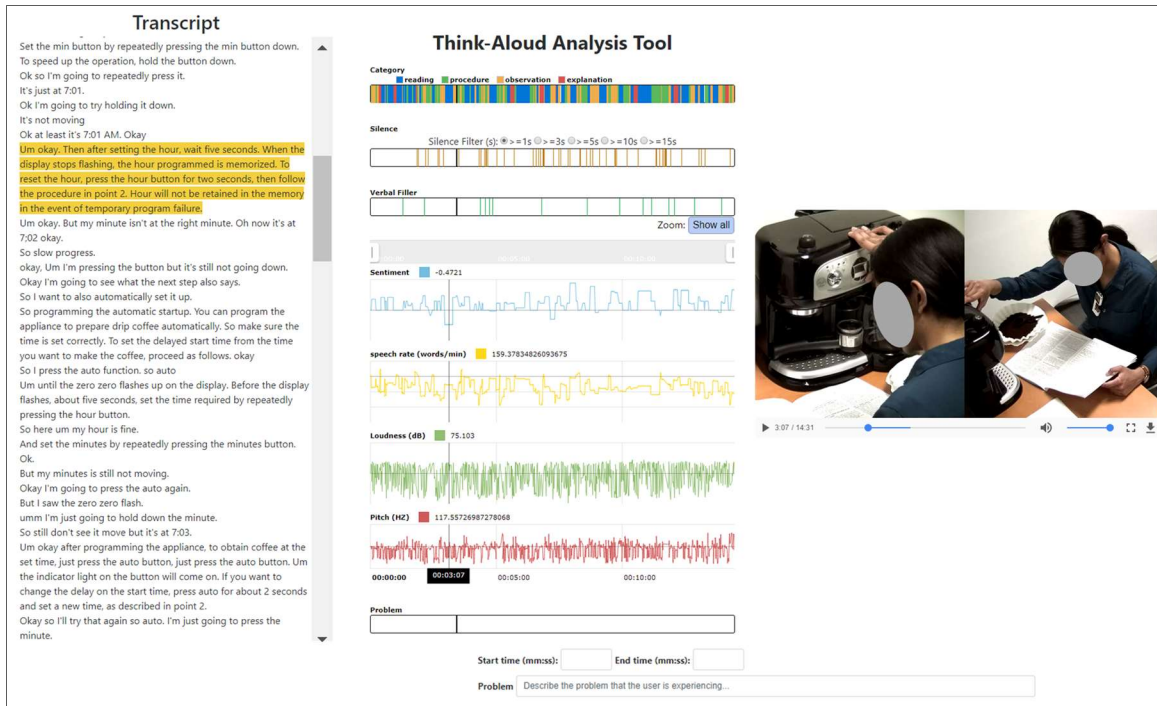


Figure 31. The think-aloud analysis tool’s interface for the *Video + Visualization* condition. The left two columns are the same as the *Audio + Visualization* condition

In all conditions, usability evaluators were given problem logging functions at the bottom of the interface. Evaluators were asked to specify the time period during which the user in the think-aloud session encountered a problem and describe the problem in plain text briefly using the logging functions in area C of the tool (Figure 25). The tool also logged when the audio or video was played. All the information was automatically saved into a log file.

6.2.5 Procedure

Prior to the start of the study, the study facilitator informed evaluators that they would use a tool to identify usability problems. The facilitator introduced the tool's functions, how to use it, and then gave each evaluator a few minutes to familiarize themselves with the tool. Because think-aloud sessions varied in length, it would be hard to allocate a fixed amount of time for analysis of different think-aloud sessions. Instead, evaluators were allocated 1.5 times the length of a session to spend on their analysis. When the evaluator finished analyzing a session or if allocated time was up, the evaluator was asked to proceed to their next session. With this design, the study lasted about 2 hours in total. Each evaluator was compensated with \$40.

6.3 Analysis and Results

6.3.1 Number of Labels per Verbalization Category

The number of times that the four verbalization categories were used as labels in all the recorded think-aloud sessions were quantified. Table 8 displays this information for each device and website separately and together. The results appeared to be similar to that of Study 2 (see Table 4), despite adding digital systems as testing objects and the new pool of participants who took part in the study.

The labels used for the four verbalizations appeared in similar proportions across the devices and websites: 1) roughly 60% of the verbalizations were about users reading contents (*Reading*) or describing their actions (*Procedure*); 2) the *Observation* category was the most popular single category and slightly over one third of verbalizations were given the *Observation* label; 3) the *Explanation* category was the least popular category and appeared significantly less than all the

other three categories. The result of the Mauchly's test shows that the sphericity assumption was violated ($\chi^2(5) = 43.29, p = .000$), therefore the degrees of freedom was corrected. A repeated-measures ANOVA test with a Greenhouse-Geisser correction found significant differences between four categories ($F(1.20, 17.99) = 75.03, p = .000, \eta_p^2 = .83$). Post-hoc pairwise comparisons show that: 1) the *Observation* category appeared significantly more than the *Reading* category ($p = .000$), the *Procedure* category ($p = .000$), and the *Explanation* category ($p = .000$); 2) the *Exploration* category also appeared significantly less than the *Reading*, the *Procedure* or the *Observation* category ($p = .000$); 3) no significant difference between the *Procedure* and the *Reading* categories ($p = .196$).

Table 9. The percentage of audio segments labeled with each verbalization category for each testing object.

Device or website	Verbalization category			
	Reading	Procedure	Observation	Explanation
Coffee machine	28.4%	28.4%	35.4%	7.8%
Universal remote	29.4%	29.6%	36.9%	4.1%
Science & tech museum	23.7%	30.7%	37.2%	8.4%
History museum	23.0%	35.7%	37.2%	4.1%
All together	26.1%	31.1%	36.7%	6.1%

6.3.2 Problems Identified by Usability Evaluators

In total, usability evaluators identified 418 problems. Two researchers validated each problem that evaluators had logged using the analysis tool, by checking their problem description and listening to (or watching) the corresponding audio (or video) segment. Any disagreements about the correctness of logged problems were discussed and resolved. Of these problems, 33 were assessed to be invalid because these problems either 1) missed the starting or ending timestamp, which made it impossible to know when evaluators thought that users were encountering problems; or 2) the problem descriptions provided by evaluators could not be inferred from the corresponding audio or video segments. With these problems removed, a total of 385 problems were considered in all the subsequent analyses.

The average number of problems identified per evaluator for each physical device or digital website was as follows: coffee machine ($M=5.9$, $SD=2.3$), universal remote ($M=5.1$, $SD=3.8$), science and tech museum ($M=6.6$, $SD=5.3$), and history museum ($M=6.5$, $SD=3.4$). The result of the Mauchly's test shows that the sphericity assumption was not violated ($\chi^2(5) = 8.59, p = .127$). A repeated-measures ANOVA test found no significant difference ($F(3, 45) = 1.42, p = .25, \eta_p^2 = .09$).

6.3.3 Verbalization Categories and the Identified Problems

I followed the same procedure as described in Study 2 to analyze the relationship between verbalization categories and the identified problems. I computed the **precision**, **recall**, and **F-Measure** of each verbalization category in locating problems. Results (Figure 32) show that the *Observation* category was most likely associated with problems. The *Explanation* category was still least likely associated with problems. The general trend shown in Figure 32 is consistent with the trend shown in Study 2 (Figure 26).

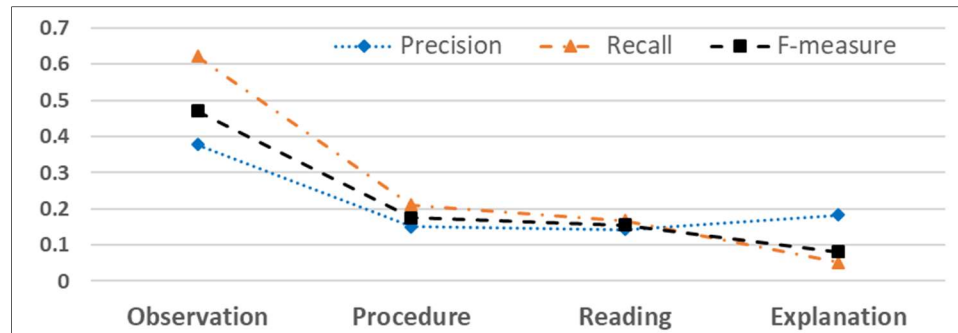


Figure 32. Precision, recall, and F-measure of each verbalization category in identifying problems.

I further computed the precision, recall, and F-measure of each *verbalization pair* in identifying usability problems (Figure 33). The results reveal that the pairs with the highest precision and recall all contained the *Observation* category. Particularly, pairs of *Observation* and *Procedure* (P-O or O-P) and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with

problems. Pairs of *Observation* and *Explanation* (E-O or O-E) had relatively high precision. The implication here is that a large number of usability problems can be detected simply by focusing on the *Observation* category as the *Observation* category has higher precision, recall, and F-measure than any given pair.

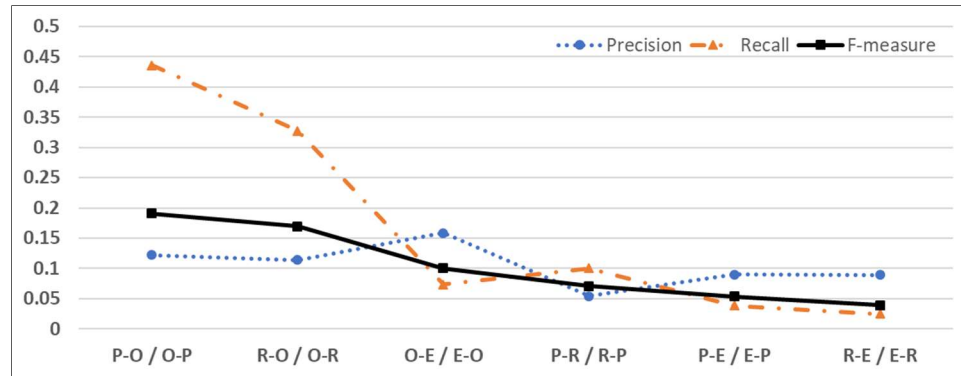


Figure 33. Precision, recall, and F-measure of each verbalization category pair in identifying problems.

6.3.4 Physical Devices vs. Digital Systems

I analyzed whether there were differences between *physical devices* and *digital systems* on how verbalization patterns may be related to usability problems, by grouping the verbalizations for the two *physical devices* and two *websites* in the analysis. Table 10 shows the results. Similar trends to that in Study 2 (Table 4) in each verbalization category's frequency of occurrence were observed in this study. Particularly, when considering physical devices and websites separately, *Observation* was still the most frequently occurring category, whereas *Explanation* was the least.

Table 10. The percentage of audio segments labeled with each verbalization category for physical and digital products respectively.

Physical or digital products	Verbalization category			
	Reading	Procedure	Observation	Explanation
Physical devices	28.8%	28.9%	36.0%	6.3%
Digital websites	23.3%	33.2%	37.2%	6.2%
All together	26.1%	31.1%	36.7%	6.1%

To understand if there was any difference in the number of reported problems for physical and digital products, I computed the number of problems that the usability evaluators identified for physical devices and digital websites respectively. The average number of problems identified per evaluator for the physical devices was 10.9 ($SD=5.8$). For digital websites, it was 13.1 ($SD=8.1$). A paired samples t-test shows that there was no significant difference between the number problems identified for the physical devices and the digital devices ($t(15) = -1.59, p = .133$).

I further examined whether physical devices and digital websites affect how verbalization categories relate to the problems by computing the precision, recall, and F-measure of each verbalization category in identifying problems for the physical devices and the digital websites separately (Table 11). Results show a similar trend to that of the Study 2. Specifically, the *Observation* category was the most relevant category to usability problems while the *Explanation* category was the least relevant category to usability problems. However, there was a difference between the *Procedure* category and the *Reading* category. Compared to the physical devices, the *Procedure* category was more relevant to usability problems than the *Reading* category for the digital websites.

Table 11. Precision, recall, and F-measure of each verbalization category in identifying problems for physical devices vs. digital websites.

Category	Precision		Recall		F-measure	
	Physical devices	Digital websites	Physical devices	Digital websites	Physical devices	Digital websites
Observation	0.36	0.40	0.61	0.57	0.45	0.47
Procedure	0.11	0.19	0.14	0.25	0.12	0.22
Reading	0.14	0.15	0.19	0.13	0.16	0.14
Explanation	0.17	0.19	0.05	0.05	0.08	0.07

I also computed the precision, recall, and F-measure of each verbalization pair in identifying usability problems for the physical devices and the digital websites separately (Table 12). Results show a similar trend that pairs of *Observation* and *Procedure* (P-O or O-P) and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with problems. One difference is that pairs of *Observation* and *Reading* (R-O or O-R) were more relevant to problems for physical

devices while pairs of *Observation* and *Procedure* (P-O or O-P) were more relevant to problems for digital websites.

Table 12. Precision, recall, and F-measure of each verbalization category pair in identifying problems for physical devices vs. digital websites.

Category pair	Precision		Recall		F-measure	
	Physical devices	Digital websites	Physical devices	Digital websites	Physical devices	Digital websites
P-O / O-P	0.09	0.15	0.35	0.50	0.15	0.23
R-O / O-R	0.11	0.12	0.43	0.25	0.18	0.16
O-E / E-O	0.18	0.14	0.09	0.06	0.12	0.08
P-R / R-P	0.03	0.08	0.08	0.12	0.05	0.09
P-E / E-P	0.04	0.15	0.02	0.05	0.03	0.08
R-E / E-R	0.08	0.10	0.03	0.02	0.05	0.03

6.3.5 Audio vs. Video

To analyze the effect of the *modality* that evaluators had for analyzing the think-aloud sessions, I first computed the number of problems that usability evaluators identified when they were given the audio recording only and when they were given the video recording as well. Results show that evaluators found on average 12.5 ($SD=7.54$) problems when they had access to only the video recording and 11.6 ($SD=5.98$) problems when they had access to the audio recording also. A paired samples t-test shows that the difference for the number of problems identified when evaluators had access to the audio or video modality was not statistically significant ($t(15)=-0.906, p=.379$).

I then conducted the same analysis to examine if the modality affects the verbalization categories and category pairs associated to the problems identified by the evaluators by computing the precision, recall, and F-measure of each verbalization category in identifying problems for the physical devices and the digital websites respectively. Table 12 shows the results. The three measures of how each verbalization category relates to problems are consistent when the evaluators had access to the audio or video modality of the think-aloud sessions. Regardless of the modality, the *Observation* category was again the most relevant to the usability problems in terms of the three measures while the *Explanation* category was the least relevant.

Table 13. Precision, recall, and F-measure of each verbalization category in identifying problems when evaluators had access to the audio or video modality of the sessions.

Category	Precision		Recall		F-measure	
	audio	video	audio	video	audio	video
Observation	0.42	0.34	0.57	0.62	0.48	0.44
Procedure	0.18	0.13	0.20	0.20	0.19	0.16
Reading	0.17	0.11	0.19	0.13	0.18	0.12
Explanation	0.18	0.18	0.05	0.05	0.07	0.08

Table 13 shows the measures of each verbalization category pair in identifying problems when evaluators had access to the audio or video modality of the think-aloud sessions. The general trend for each modality is consistent with pairs of *Observation* and *Procedure* (P-O or O-P), and pairs of *Observation* and *Reading* (R-O or O-R) were most likely associated with problems.

Table 14. Precision, recall, and F-measure of each verbalization category pair in identifying problems when evaluators had access to the audio or video modality of the sessions.

Category pair	Precision		Recall		F-measure	
	audio	video	audio	video	audio	video
P-O / O-P	0.16	0.10	0.46	0.41	0.23	0.16
R-O / O-R	0.10	0.13	0.32	0.33	0.16	0.18
O-E / E-O	0.13	0.19	0.07	0.08	0.09	0.11
P-R / R-P	0.05	0.06	0.10	0.10	0.06	0.08
P-E / E-P	0.11	0.08	0.04	0.04	0.06	0.05
R-E / E-R	0.04	0.18	0.01	0.04	0.02	0.06

6.3.6 With vs. Without Visualization

To analyze the effect of visualization, I grouped the problems based on whether the evaluators had access to the visualization or not and computed the number of problems that they identified with and without accessing to the visualization. The results show that evaluators, on average, identified 10.8 ($SD=5.4$) problems with the visualization, and 13.3 ($SD=8.3$) problems without the visualization. A paired sample t-test shows that the difference was not statistically significant ($t(15) = -1.888, p = .078$).

Table 15. Precision, recall, and F-measure of each verbalization category in identified problems when evaluators worked with or without visualization.

Category	Precision		Recall		F-measure	
	with	without	with	without	with	without
Observation	0.41	0.34	0.59	0.60	0.48	0.43
Procedure	0.18	0.12	0.22	0.18	0.20	0.14
Reading	0.16	0.12	0.15	0.17	0.16	0.14
Explanation	0.20	0.16	0.05	0.05	0.07	0.08

I computed the precision, recall, and F-measure to examine if the visualization affects how verbalization categories and category pairs relate to the identified problems. Results for verbalization categories and category pairs are shown in Table 15 and Table 16 respectively. The general trend of the measures is consistent when evaluators worked with or without visualizations.

Table 16. Precision, recall, and F-measure of each verbalization category pair in identified problems when evaluators worked with or without visualization.

Category pair	Precision		Recall		F-measure	
	with	without	with	without	with	without
P-O / O-P	0.14	0.09	0.46	0.39	0.22	0.16
R-O / O-R	0.13	0.10	0.31	0.37	0.18	0.16
O-E / E-O	0.17	0.14	0.06	0.09	0.09	0.11
P-R / R-P	0.07	0.04	0.10	0.10	0.08	0.06
P-E / E-P	0.10	0.07	0.04	0.04	0.05	0.05
R-E / E-R	0.14	0.04	0.03	0.01	0.05	0.02

6.3.7 What Users Talked About When They Encountered Problems?

To further examine the relationship between verbalization categories and the usability problems at the word level, I calculated the most frequently uttered words that users verbalized when encountering problems. I removed the stop words (e.g., pronouns, articles, common verbs such as be) and plotted the top 30 most frequently verbalized words when users encounter problems. Figure 34 shows the result.

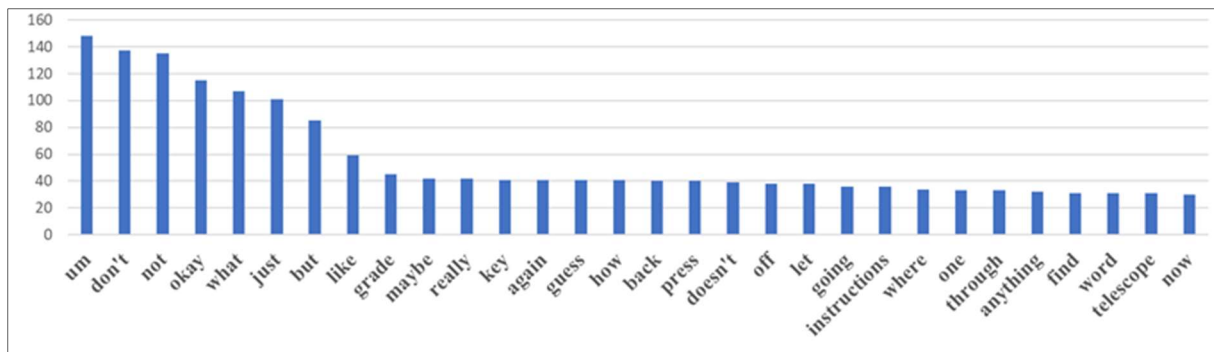


Figure 34. Most frequently verbalized words when users encountered problems.

Based on the results, the most frequently verbalized words consisted of 1) verbal fillers, such as *um*; 2) negations, such as *don't* and *not*; *doesn't*; 3) words expressing uncertainty, such as *maybe* and *guess*; 4) words signaling repetitive effort, such as *again* and *back*; 5) words used to raise questions, such as *what*, *how* and *where*; 6) nouns related to the tasks or the test products, such as *grade*, *key*, and *telescope*; 7) verbs related to the tasks or the test products, such as *press*, and *find*.

I further analyzed the verbalizations that were associated with each problem to better understand the utility of these frequently occurring words: 1) *how often did think-aloud users use verbal fillers (e.g., um)?* 2) *how often did think-aloud users use negation (e.g., not, don't, doesn't)?* 3) *how often did think-aloud users use uncertain words (e.g., maybe, guess)?* 4) *how often did think-aloud users use words suggesting repetition (e.g., again, back)?* 5) *how often did think-aloud users ask themselves questions (e.g., what, how)?* It is worth mentioning that the word *Okay* was not included as a verbal filler in the analysis since it can also be used for confirmation.

The results show that out of the 385 problems, think-aloud users used: 1) *negation* words in 266 (69%) problems; 2) *filler* words in 148 (38%) problems; 3) words showing *uncertainty* in 95 (25%) problems; 4) words that *raised questions* in 94 (24%) problems (e.g., “*did I miss anything?*”); 5) words showing *repetitive effort* in 56 (15%) problems. It is worth mentioning that uncertainty was not always expressed through a single signaling word (e.g., *maybe*, *guess*). It was sometimes expressed through *their verbalized actions* (e.g., “*I'm just clicking some random links on this*

page”). Furthermore, I also noticed that in 41 (11%) problems, users experienced *Aha! Moments*, which were the moments when they suddenly came to an understanding of something that they had previously misunderstood or could not understand (e.g., “*oh, I thought they meant the power key.*”). This phenomenon was previously articulated by evaluators in Study 1 as well. Usability evaluators also identified five problems in which users articulated *suggestions* (e.g., “*it would be better if I could filter through them to choose grade*”).

6.3.8 Speech Features and Identified Usability Problems (i.e., How Did Users Verbalize When Experiencing Problems?)

The interviews in the previous two studies (i.e., Study 1 and Study 2) showed that evaluators considered all the verbalization and speech features useful as cues for identifying problems. Although the evaluators were mostly positive toward using the verbalization categories, silence, filler words and sentiment for problem identification, their thoughts on speech rate, loudness and pitch were mixed. Motivated by this finding, I further computed the same three measures (i.e., precision, recall, F-measure) for each speech feature to quantitatively understand whether and how the speech features were related to the usability problems.

In this analysis, I considered a speech feature’s value to be abnormal (i.e., high or low) if it was greater or less than two standard deviations away from the feature’s average value in the whole audio recording. Figure 35 shows the precision, recall, and F-measure of each speech feature in identifying problems. While high and low pitch has high recall values (i.e., 0.86 and 0.80 respectively), the low speech rate has a high precision value (0.70). The implication is that if evaluators examined all the verbalization segments with abnormal pitch values, they would have a high chance to locate a high percentage of all the usability problems due to the high recall values. If evaluators examined all the verbalization segments with a low speech rate value, they would have a high success rate in finding a usability problem due to the high precision value. However, at the same time, it is important to mention that there is no single voice feature that has both high precision and recall. The implication is that usability evaluators should not just rely on any single voice feature if they would like to identify as many usability problems as possible. These features

should be used together with other features, such as the verbalization categories, sentiments, negations, filler words, and words for asking questions, expressing uncertainty, or signaling repetition.

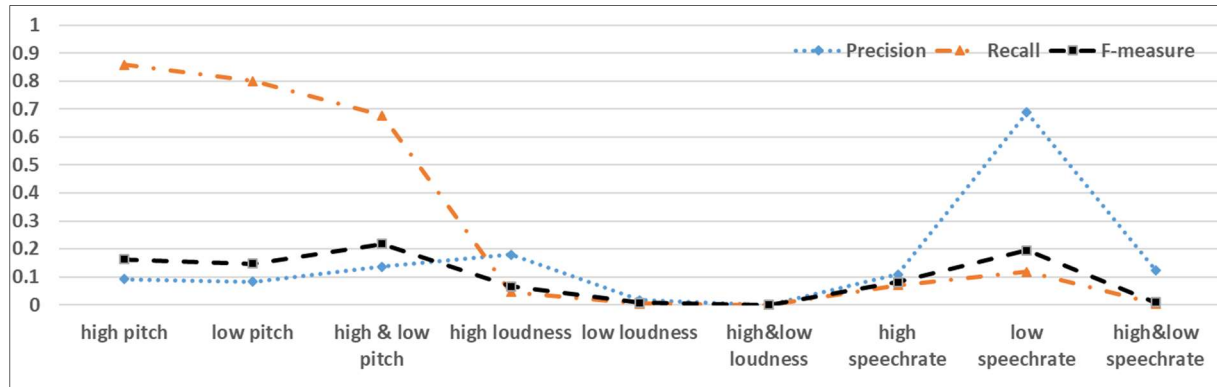


Figure 35. Precision, recall, and F-measure of each speech feature in identifying problems.

6.4 Discussion

6.4.1 Physical Devices vs. Digital Systems

The findings of this study show that the verbalization categories appeared in similar proportions for physical devices and digital systems. For example, regardless of the test product, *Observation* was the most frequently occurring category while *Explanation* was the least. The frequency of the *Reading* and *Procedure* category were also very similar.

Additionally, the verbalization patterns that cues to problems were also very similar. However, one subtle difference was that there was slightly more of the *Reading* and less of the *Procedure* category when using physical devices (Table 11), perhaps because users were not familiar with the devices and had a greater need to solicit help from instruction manuals which were readily available to them. In contrast, while navigating websites, users did not have a prescribed set of steps that they could refer to for completing tasks and thus less relied on reading from websites but more exploration.

6.4.2 Audio vs. Video Modality Available to Evaluators

The analysis of this study also shows that usability evaluators identified roughly the same number of problems when they had access to either *audio* or *video* modality of think-aloud recordings. Furthermore, the verbalization categories and category pairs that are related to usability problems are similar when evaluators had access to either the audio or video modality of the recorded think-aloud sessions.

I further analyzed the problem descriptions that evaluators provided when they identified these problems to understand the types of problems that they identified when having access to a different modality of the think-aloud recordings. The result shows that evaluators found roughly the same types of usability problems when they had access to different modalities in the study. During the interviews with the evaluators, many expressed that even when they did not have the video stream, the richness of the sounds in the audio stream helped evaluators imagine what users were experiencing (“*Yes, [without the video stream] I can’t see their faces or their interactions with the interfaces. But I can still experience their emotions and struggles by listening to the audio.*”-s3_ev15). Evaluators also consistently agreed that the audio was useful not only because the verbalized words provided insight into users’ thought process and their feelings (based on the tone/pitch of their voice), but also because non-words uttered by users and even the noise from the surrounding environment provided valuable contextual information. For example, the sound of sighing could indicate that a user is frustrated. Frequent page flipping sounds could indicate that an instruction manual was poorly designed, subjecting users to constantly revisiting pages. Mechanical sounds generated by operations on devices, such as the clicking of a button, can help evaluators understand the fluency of a user’s actions.

It is worth noting that evaluators who had access to the video modality, however, did provide evidence from the visual channel that was unavailable from the audio channel to support problems. In other words, evaluators who had access to the audio modality could describe the same problem using different evidence than the evaluators who had access to the video modality. For example, the evaluators who had access to the audio modality inferred the mismatch between the product

instructions and the actual device from what the users said for the universal remote control. In contrast, the evaluators who had access to the video modality pointed out that some labels on the device did not match with the instructions, which required an ability to observe the actual device. Similarly, the evaluators who had access to the video modality pointed out the issue of the lack of images on the searching result page, another one that required the ability to observe the website, to illustrate the content presentation issue of the science & technology museum website.

Thus, this difference in using different types of evidence to support their reported usability problems suggests that although evaluators found that having access to the audio stream only when analyzing think-aloud sessions was sufficient most of the time, they still found the video stream to be useful. Being able to see a user's face could be helpful because facial expressions could reflect their mood. But because some users kept a neutral face throughout the entire think-aloud session, seeing the user's face was not always useful. How facial expressions could be used to identify usability problems remains to be explored. Moreover, evaluators pointed out that a user's body language could also signal problems. For example, one evaluator found that a user had a tendency to scratch his head whenever he encountered problems. However, it is unclear whether body language is a reliable and consistent cue to locate usability problems.

Additionally, evaluators also felt that there were particular times when having access to a video recording would have been useful. For example, a video stream would be valuable when users become silent since evaluators sometimes had a hard time determining whether users were stuck or just waiting for something to happen. This finding suggests that it might be a good idea to draw attention to the video stream of a recording when the think-aloud users become silent or maybe slightly before they fall into silence. Additionally, evaluators pointed out that video could be important when think-aloud users verbalized their actions using demonstratives (i.e., this, that, these, those) or adverbs of place (e.g., here, there). For example, verbalizations, such as "*I'm going to hold **this** button and **this** button*" or "*I'm clicking the link **here**,*" can be hard to understand without seeing what users are referring to. On the other hand, evaluators might not want to constantly monitor the video when having access to it ("*without video, I can concentrate better on listening. If needed, I'll look at the video*"-s3_ev12). Thus, one interesting question would be to

help evaluators figure out what users are implying when they verbalize vague statements and to highlight moments in the videos that should be given attention, to reduce the need to monitor the video stream constantly.

6.4.3 Visualization of Verbalization and Speech Features

The think-aloud analysis tool provided visualizations of verbalization categories and six speech features to evaluators in the *Audio + Visualization* and *Video + Visualization* conditions, which was a novel feature that had not been explored in the literature. These two conditions were compared with two baseline conditions, which were the *Audio only* and the *Video only* conditions. The results of this study showed that the number of problems that evaluators identified with access to the visualizations was not significantly more than having no access to them, and the patterns in verbalization and speech features relating to problems were also similar when evaluators had or did not have access to the visualizations.

One possible reason might be that the way these features were presented in the tool might have overwhelmed evaluators. This is evident in evaluators' feedback. Particularly, one evaluator reported that she almost completely ignored the visualizations because the interface was "too busy." This raises an interesting challenge for future exploration: *how to visualize the verbalization categories and speech features to maximize their utility to usability evaluators?*

6.4.4 Verbalization Category Proportions

When conducting think-aloud sessions, I followed Ericsson and Simon's three guidelines: use neutral instructions, allow participants to practice thinking aloud, and no probe or intervene during think-aloud sessions except to remind participants to keep talking if they fall into silence for a long time [32]. One study that examined users' verbalizations when following these guidelines was conducted by Zhao *et al.* [110]. In their study, authors analyzed users' verbalizations in think-aloud sessions that were conducted under two conditions: the *classic instruction* condition and the *explicit instruction* condition. The *classic instruction* condition strictly followed all three guidelines advocated by Ericsson and Simon. In contrast, the *explicit instruction* condition was the

same as the *classic instruction* condition, except that it included an explicit instruction requesting participants to report both the explanations and verbalizations that are relevant to understanding the user experience.

In their study, users' verbalizations were categorized into five categories: *procedural description*, *positive experience*, *negative experience*, *expectation*, and *explanation*. Based on the definitions of these categories, the relationship between these categories and the four categories that were used in the study was as follows: *procedural description* is equivalent to the combination of the *Reading* and the *Procedure* categories; *positive experience*, *negative experience* and *expectation* together are equivalent to the *Observation* category; *explanation* was equivalent to the *Explanation* category. As a result, I combined the *Reading* and the *Procedure* categories into one category and computed the average proportion of the verbalization categories in this study (i.e., Study 3) and the previous study (i.e., Study 2). Table 17 shows the result.

Table 17. Verbalization category proportion.

Studies	Verbalization category		
	Reading, Procedure	Observation	Explanation
Our studies (i.e., Study 2 and Study 3)	56.3%	37.6%	5.9%
Classic Instruction condition in Zhao et al. [110]	70.3%	20.1%	9.6%
Explicit Instruction condition in Zhao et al. [110]	49.9%	33.8%	16.3%

Based on the result, it is evident from the result in Table 16 that verbalizations of the *Observation* and the *Explanation* categories exist are present even when following the guidelines proposed by Ericsson and Simon's guidelines. In other words, users do verbalize their comments, feelings, and rationales (labeled as the *Observation* and the *Explanation* categories) even when users were not explicitly instructed to do so. Second, both our studies and the two conditions in Zhao *et al.*'s study found that the majority of the verbalizations fall into sequences of the *Reading* and the *Procedure* categories. Third, both our studies and the explicit instruction condition had less amount of the *Reading* and the *Procedure* categories compared to the classic instruction condition. For the explicit instruction condition, this was because the explicit instruction was given, which was evident from the significantly higher number of occurrences of the *Explanation* category. I

reflected on how I conducted think-aloud sessions and how the process might have differed from that of Ericsson and Simon to explain this marked increase in the amount of the *Observation* category (and similarly, the decrease in the amount of the *Reading* and the *Procedure* categories). I noticed that although I followed the three guidelines proposed by Ericsson and Simon [32], I also showed the participants a one-minute demo video of a think-aloud session being carried out by an actor, which is offered online by the Nielsen and Norman group [75]. In this one-minute demo video, the participant verbalized her comments and feelings about a test website in addition to describing her actions and what she sees on the website. This demo video might have implicitly influenced the participants to verbalize their comments and feelings, in their attempts to mimic the actor in the demo.

6.4.5 Evaluator Effect

Previous studies reported that evaluators might find different sets of usability problems, even when they analyze the same usability test sessions (e.g., [43,44]). I further analyzed this study's data to see if evaluators who analyzed the same think-aloud session would agree on the verbalization segments that were linked to problems. To measure the agreement between two evaluators, I computed the *any-two agreement* measure using the following equation: $\frac{P_i \cap P_j}{P_i \cup P_j}$ (P_i and P_j are the sets of problems identified by two evaluators i and j) [43] for each think-aloud session that was evaluated by two evaluators. I then computed the average any-two agreement for all test products (the first column in Table 18) and that for each test product separately (the second to the last columns in Table 18).

Table 18. the average any-two agreement between evaluators.

All test products together	Coffee machine	Universal Remote	History museum website	Science & technology museum website
0.80	0.76	0.88	0.82	0.76

The values of the average any-two agreement measure in Table 18 show that the evaluators in our study had a reasonably high agreement. The patterns between think-aloud verbalizations and the

usability problems that I identified in this research were based on the analysis of the joint problems identified by participants ($P_i \cup P_j$). The relatively high agreement between our evaluators suggests that the identified patterns would be largely applicable to each individual evaluator although they might disagree if a verbalization segment indicates a problem sometimes.

The average any-two agreement in our study was similar to the average any-two agreement reported in other studies (e.g., 0.71 in [110]) and was higher compared to other studies in the literature. For example, the average any-two agreement was 31% for moderated sessions, wherein a moderator presented and probed the user, and was 30% for unmoderated sessions, wherein no moderator was present [44]. Many factors could contribute to the differences. One factor was the amount of time that was allocated to evaluators to analyze the sessions. Evaluators in our study were allocated 1.5 times the length of a think-aloud session to spend on their analysis.

In contrast, the evaluators in Hertzum *et al.*'s study [43] spent on average 22 hours to analyze the sessions, which were on average 33 minutes. Therefore, in our study, evaluators may be more likely to focus on the more significant issues, leading to a higher agreement between evaluators. In fact, the evaluators in Hertzum *et al.*'s study [43] also had a much higher any-two agreement for critical problems (53% for moderated sessions, and 69% for unmoderated sessions), which was more closer to our measures. However, there are other factors that may have resulted in the difference between the any-two agreement measures. For example, our study used four test products, which consisted of two physical devices and two digital websites, while their test product was one digital website. Further, in our study, two evaluators examined each think-aloud session while their evaluation had nine or ten. The background and experience of think-aloud participants and evaluators may have also contributed to the differences.

6.5 Summary

Through a series of three studies (Chapter 4, Chapter 5, and Chapter 6), I systematically studied the relationship between verbalization and speech features and usability problems in concurrent think-aloud sessions and identified and validated the patterns in users' verbalization and speech

features that tend to occur when users encounter usability problems. The findings from these three studies repeatedly show that certain patterns of verbalization and speech features act as telltale signs of usability problems. Segments labeled as the *Observation* category were most likely associated with usability problems. Segments labeled as the *Procedure* category that also contain a description of repeated actions were likely associated with usability problems. Segments labeled as the *Reading* category that last for a long period of time were also likely associated with usability problems. On the contrary, segments labeled as the *Explanation* category were relatively rare and did not have a clear relationship with usability problems.

The findings further show that evaluators often identified problems using combinations of verbalization categories since category combinations were helpful in providing contextual information as to why users were encountering problems. Furthermore, pairs of verbalization categories that contained the *Observation* category were generally more likely associated with problems than those without the *Observation* category.

The findings demonstrate that although the *Observation* category was most indicative of usability problems of all four categories, the F-measure of using the *Observation* category to locate usability problems were around 0.5. To increase the chance of locating a problem, *sentiment* and *speech features* should be considered in conjunction with the category information. For example, when experiencing problems, users tended to use *negations*, *verbal fillers*, words indicating *uncertainty*, *repetitions*, or *questions*. Therefore, the sentiment of these verbalizations was often *negative*. Furthermore, users tended to verbalize their thought units in *high or low pitch* or with *low speech rate* but rarely changed the loudness of their voices when experiencing problems.

The verbalization and speech patterns that tend to occur when users encounter problems are largely generalizable to three factors: the types of test products (i.e., physical devices vs. digital systems), the modality used to record the think-aloud sessions that evaluators were provided with (i.e., audio vs. video recording), and access to a visualization of the verbalizations. The implication is that the same set of verbalization and speech patterns can be used to identify problems that users were experiencing when thinking aloud regardless of whether a physical device or a digital system was

used. Usability evaluators can rely on verbalization and speech features alone to identify problems by and large, although certain cues in video streams have additive values to their analysis, such as facial expression and body language. However, whether these visual cues are consistent across users for locating problems remains to be examined. Moreover, the video stream of a think-aloud session can be informative when the think-aloud user remains silent or frequently uses demonstratives (e.g., this, that) or adverbs of place (e.g., here, there), which makes it difficult to infer what the user is referring to from the audio stream alone. As a result, in such situations, it would be preferable to draw evaluators' attention to the video stream. This study (i.e., Study 3) also reveals that visualizations of verbalizations as provided in the studies did not affect the number of problems identified or the verbalization and speech patterns that were associated with problems.

One logical next step is to *design systems* that leverage these verbalization and speech patterns that tend to occur when users encounter problems to *automatically detect when in a recorded think-aloud session users encounter problems*. Such automatically inferred usability problem encounters could then be used to draw usability evaluator's attention to parts of the session that are more likely to reveal problems, which could potentially improve their analysis efficiency and experience. In the next chapter, I will describe the computational methods that automatically detect usability problem encounters based on the patterns in users' verbalization and speech features that tend to occur when users encounter problems.

Chapter 7 Automatic Detection of Usability Problem Encounters

I have identified and validated the users' verbalization and speech patterns that tend to occur when they encounter usability problems in concurrent think-aloud sessions via a series of three studies, each addressing the limitations of the previous one. Informed by the findings, I hypothesize that these subtle patterns can be used to detect the encounters of usability problems automatically because these patterns would allow computational methods to capture key characteristics of usability problem encounters better than the generic text or speech features that are agnostic to these hidden honest signals in users' verbalization and speech features.

To develop effective methods, I leverage the power of natural language process (NLP) and machine learning (ML) technologies, which have become increasingly powerful and are gradually adopted to tackle challenging problems in qualitative research [45]. For example, researchers have designed ML methods to automate or semi-automate qualitative coding [63,100,101], to detect potential disagreements in qualitative coding between coders [17], and to generate human-understandable explanation that reveals AI's internal states [28]. Inspired by this line of research, in this work, I took the first step to *design and evaluate computational methods to automatically detect the encounters of usability problems in think-aloud sessions*. Specifically, this chapter answers the third research question of this dissertation:

RQ3: Can the subtle verbalization and speech patterns be used to detect the encounters of usability problems automatically?

7.1 Research Questions

I aim to answer to the following sub-questions together and use the answers together to better answer the third research question of this dissertation (RQ3):

- RQ3-1: Can *users' verbalizations* during think-aloud sessions be used to detect usability problem encounters?

In other words, can we create a model based on just what users say during think-aloud sessions to detect when they encounter problems?

- RQ3-2: Can *the verbalization and speech patterns that tend to occur when users encountered problems* be used to improve the detection of the usability problem encounters?

In other words, can we add a model based on how users speak during think aloud sessions to make this detection of when they encounter problems even better?

Furthermore, as UX practitioners typically work for a company on a specific product to improve its user experience, it can be beneficial *to build ML models for a product that can detect the encounters of usability problems by new test users* because UX practitioners can leverage the detection results to speed up their analysis for these new test users. Similarly, it is not uncommon for companies to maintain a pool of volunteer testers whom they may contact for usability testing to save the recruitment cost. Therefore, it can be beneficial *to build an ML model for a user (e.g., a volunteer in the company's volunteer pool) that can detect the problems that the user may encounter when using a new product* because UX practitioners can use the detection results to help them identify usability problems with the new product. As a result, I also sought to answer the following two research questions:

- RQ3-3: Can an ML model be built for *a product* using its existing users' think-aloud data to detect the usability problem encounters by a *new* user when using the product?
- RQ3-4: Can an ML model be built for *a user* using the think-aloud data of the products with which the user has interacted to detect the usability problem encounters by the user when using a *new* product?

7.2 Automatic Detection of Usability Problem Encounters

To answer the research questions, I first needed to curate a dataset of recorded think-aloud sessions. To do so, I used the dataset of the recorded think-aloud sessions that were collected in

Study 3. The usability problem encounters were labeled by UX researchers as ground truth and the verbalization and speech features were labeled or computed as input features to train ML models. To better understand how different ML methods fared in detecting the usability problem encounters, I implemented and evaluated a wide range of ML methods.

7.2.1 Think-Aloud Dataset

I used the dataset of recorded think-aloud sessions that were collected in Study 3 (Chapter 6 generalization study), which was the largest dataset I collected (i.e., 64 recorded think-aloud sessions) and contained data of users using both physical and digital products, to build and evaluate the computational models for identifying usability problems.

7.2.2 Ground truth Labeling

The think-aloud sessions were first manually transcribed into text. Then, two coders followed a similar approach used in previous work [24,30] to divide each think-aloud session recording into smaller segments to facilitate further annotation. The beginning and the end of a segment was determined by pauses between verbalizations and the verbalization content [24,30]. Each segment could include single words, but also clauses, phrases, and sentences. For each segment, two coders labeled independently whether the user experienced a problem (e.g., being frustrated, confused or experiencing a difficulty). Upon completion, they discussed to consolidate the *problem* label (i.e., 0 or 1) for each verbalization segment in the dataset, *which* was used as the *ground truth*.

7.2.3 Basic Transcript-based Feature Extraction

I computed basic text features from the transcript of each verbalization segment on the dataset. Specifically, for the transcript of each segment in each recorded session, I computed the *TF-IDF* (i.e., term frequency-inverse document frequency) feature vector using the Scikit-learn [82] and computed the *word embedding* vector using Tensorflow [1]. These vector representations and the ground truth labels of the usability problems were used together to train the ML models later.

7.2.4 Verbalization and Speech Feature Extraction

The results of the three studies have shown that users tend to verbalize the content of the *Observation category*, *negations*, *questions*, and *negative sentiment* using *abnormal pitches* and *speech rates* when they encounter usability problems in think-aloud sessions. Inspired by this finding, I aimed to evaluate whether *these verbalization and speech features* can be used to train ML models to better detect usability problems. Next, I describe how the verbalization and speech features were labeled or computed.

Verbalization Category: For each segment in the recorded think-aloud sessions, two coders independently labeled it with one of the four verbalization categories (*i.e.*, *Reading*, *Procedure*, *Observation*, and *Explanation*) [24]. Upon completion, they discussed their category labels to resolve any conflicts. In the end, each segment has a label indicating its verbalization category.

Negations: A keyword matching algorithm was designed to determine whether users verbalized a negation in a segment. The keywords were as follows: no, not, don't, doesn't, didn't, and never. Thus, each segment has a binary label to indicate whether the user used a negation.

Questions: Similar to the detection of negations, a keyword matching algorithm was designed to determine whether users asked a question in each segment. The keywords were as follows: what, which, why, how, and where. To reduce false positive detection, a heuristic that the keywords must be at the beginning of a sentence was applied to ensure the keywords were indeed used to raise questions. Thus, each segment has a binary label to indicate whether the user asked a question.

Sentiment: For each segment in the recorded think-aloud sessions, two coders independently assigned it with one of the three sentiment values (1 for positive sentiment; 0 for neutral sentiment; and -1 for negative sentiment) by referring the text transcription of the segment and listening to the corresponding audio segment if deemed necessary by the coders. Afterward, they discussed their labels to consolidate the sentiment labels for all segments.

Pitch: For the corresponding audio of each segment, I computed the user's pitch (HZ) using the speech process toolkit Praat [9].

Speech Rate: For the corresponding audio of each segment, I computed the speech rate by dividing the number of words spoken in a segment by its duration. The number of words spoken in a segment was counted based on the text transcription of the segment.

Abnormal Pitch and Speech Rate: To determine whether a segment contains *abnormal* pitch or speech rate, I computed the mean and the standard deviation of the pitch and the speech rate over the entire session recording and automatically labeled a segment as having *abnormally* high or low pitch or speech rate if any value in the segment was two standard deviations higher or lower than the mean. As a result, each segment would have two labels to indicate whether it has an *abnormally high pitch* or *abnormally low pitch* respectively and one label to indicate whether it has *abnormally low speech rate*.

7.2.5 ML Models

Recent research has shown the promise of ML in solving qualitative research problems. For example, Support Vector Machine (SVM) has shown to be effective in helping qualitative researchers code qualitative data [100,101,111] and Random Forest (RF) has demonstrated to be effective in detecting segments of question-answer from classroom conversations [8] or classifying activities in classroom discourse [96]. Therefore, I employed these two methods (i.e., SVM and RF) to detect the usability problem encounters in this work. Additionally, the convolutional neural network (CNN) and recurrent neural network (RNN) have shown to be promising on generic text classification tasks [21,56]. Thus, I also included CNN and RNN to understand how they fare in detecting usability problem encounters.

Specifically, I computed and used the TF-IDF feature vectors and the ground truths of all the segments in all recorded think-aloud sessions to train the RF and the SVM models and used the word embedding feature vectors and the ground truth of all verbalization segments to train the CNN and the RNN models. These models were referred to as the baseline. Furthermore, to evaluate

whether the six verbalization and speech features (i.e., category, sentiment, negation, question, pitch, and speech rate) can be used to improve the performance of ML models, I appended these six features to the end of the TF-IDF vector and the word embedding vector of each segment and then used these updated feature vectors and the ground truth for all verbalization segments to train the same set of ML models again. By comparing the performance of these updated models with the baseline, I was able to assess whether the six verbalization and speech features helped to improve the models' performance.

I implemented RF and SVM using the Scikit-learn library. For RF, I set the number of trees in the forest to be 200 and the max depth to be None. For the SVM model, I used the linear kernel, the L2 regularization, and the squared hinge loss function. I used the Tensorflow to implement the CNN and RNN models. The CNN model had an embedding layer followed by a convolution layer, a max pooling layer and then a softmax layer. I set the filter size to 3x4, the number of filters to 128, and the stride to be 1. The RNN model had an embedding layer followed by an LSTM with the GRU (Gated Recurrent Unit) as the RNN cell and softmax as the activation function.

7.3 Evaluation and Results

I performed the cross-validation the whole dataset (Section 7.3.1) to answer the RQ3-1 and RQ3-2. I then performed the leave-one-user-out evaluations for each product (Section 7.3.2) to answer the RQ3-3 and the leave-one-product-out evaluations for each user (Section 7.3.3) to answer the RQ3-4.

7.3.1 The Effect of the *Verbalization and Speech Features* and the *ML models* on the Detection of Usability Problem Encounters

The evaluations in this section aimed to answer the RQ3-1 and RQ3-2. I trained the ML models using the TF-IDF or the word embedding vector extracted from the transcript (i.e., users' verbalizations) as the input, which was referred to as *the transcript feature*. Furthermore, I trained the same set of ML models using both the transcript features and one of the *verbalization & speech features* (Section 7.2.4) as the input.

I performed 10-fold cross-validation to evaluate the models and used the F1-score as the overall performance measure. Figure 36 shows the result. It shows that the average F1-score of the four ML models trained on the *transcript feature* only was $.58$ ($SD=.07$). In contrast, the average F1-score of the four ML models trained with the *transcript feature* and *one additional verbalization and speech feature* was $.62$ ($SD=.02$). The increase in the performance indicates that the verbalization and speech features helped to improve the performance of the ML models. Among the four ML models, the SVM trained performed the best with the average F1-score of $.70$.

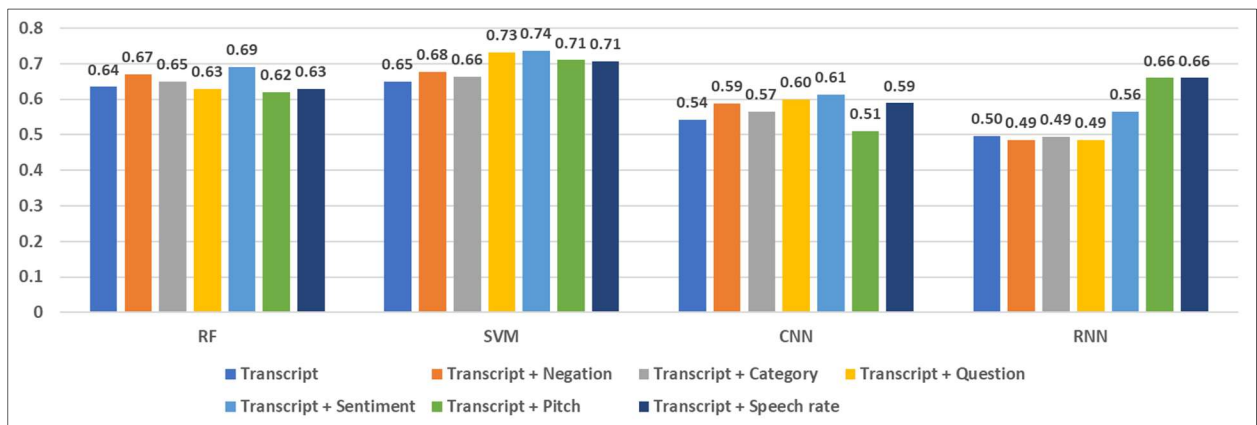


Figure 36. The average F1-score of the four types of ML models trained using *the transcript feature only* as the input or using *the transcript and one additional verbalization & speech feature* (i.e., *negation, category, question, sentiment, pitch, and speech rate*) together as the input and evaluated the models using 10-fold cross-validation on the entire dataset.

As each verbalization and speech feature was shown to improve the performance of the ML models (Figure 36), I further tested whether using all the *verbalization and speech features* together as input could improve the performance of the models even further. I trained the ML models using *all the verbalization and speech features* and *the transcript feature* (i.e., *TF-IDF or word embedding vector*) together as the input and performed a 10-fold cross validation on the entire dataset again. Figure 37 shows the precision, recall, and F1-score of the ML models trained on the transcript feature only and the transcript feature with all the verbalization and speech features together. The average F1-score of the four ML models was $.67$ ($SD=.06$), which was higher than

that of the models trained with *the transcript only* as the input feature (.58) or with *the transcript feature and any one of the verbalization & speech features* together as the input feature (.62). This finding suggests that the verbalization and speech features complement each other when used together as the input feature to train ML models. Also, the absolute difference between the precision and the recall for the RF, SVM, CNN, and RNN models when trained on the *transcript feature (i.e., TF-IDF)* and *all the verbalization & speech features together* was .16, .06, .31, and .22 respectively. This result suggests that the SVM model not only performed best in terms of F1-score but also had the most balanced precision and recall compared to the other three models (i.e., RF, CNN, and RNN).

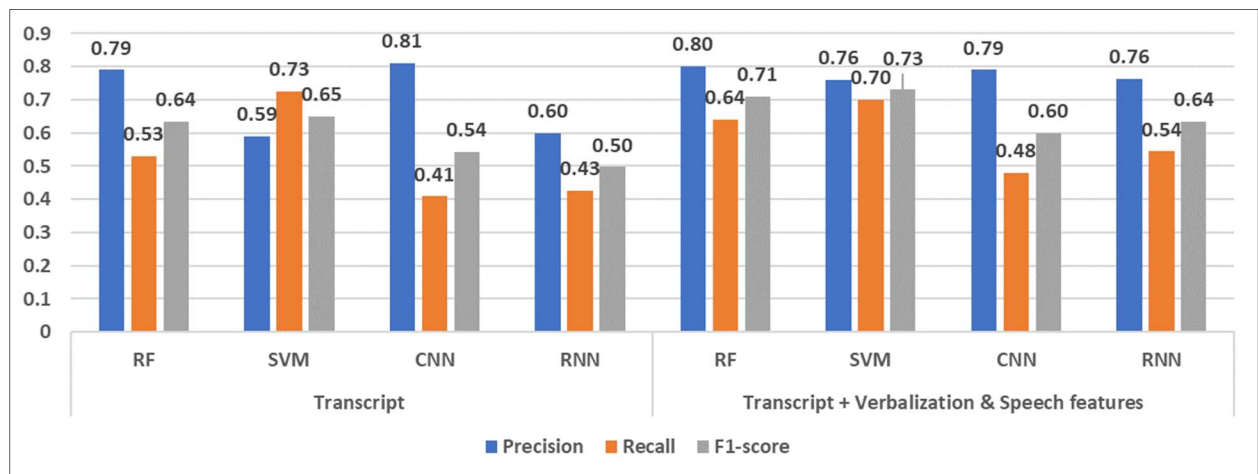


Figure 37. The precision, recall, and F1-score of the four types of ML models trained using the *transcript* feature only as the input (the left half of the figure) and using the *transcript + all verbalization & speech features* together as the input (the right half of the figure) and evaluated using 10-fold cross-validation on the entire dataset.

I used the transcript feature (i.e., TF-IDF or word embedding) as the input or part of the input to train the ML models in the evaluations so far. To further understand *whether the verbalization and speech features* alone are sufficient to train effective ML models, I used only *the six verbalization and speech features* together as the input vector to train the ML models and performed 10-fold cross-validation on the entire dataset. Figure 38 shows the precision, recall, and F1-score of the

ML models. The performance was comparative to the same models trained using both the transcript feature and the verbalization and speech features together as the input (Figure 37 right). The results also show that the SVM model performed best in terms of F1-score (.75) and had the most balanced precision and recall (i.e., the least difference between the precision and recall measures) among the four ML models.

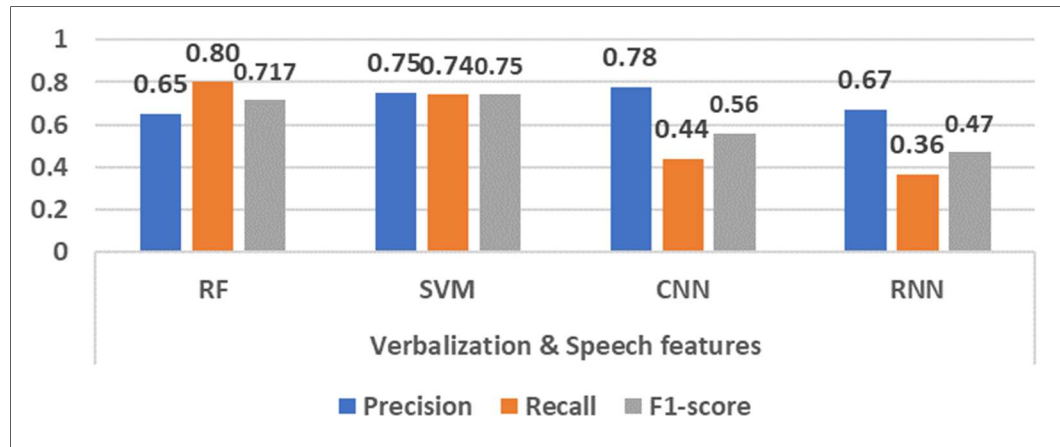


Figure 38. The average precision, recall, and F1-score of the four types of ML models trained using only the *verbalization & speech features* as the input and evaluated using 10-fold cross-validation on the entire dataset.

To further understand how each individual verbalization and speech feature contributed to the model's precision, recall, and F1-score, I used each verbalization or speech feature as input respectively to train an SVM model, the best performed model based on the evaluations so far, and performed 10-fold cross-validation on the dataset. Figure 39 shows the results. The results demonstrate that each verbalization or speech feature had different precision and recall when locating usability problems. For example, the *sentiment* and the *negation* features had a relatively higher precision in locating usability problems. On the other hand, the *category*, *pitch*, and *speech rate* features had a relatively higher recall in locating the usability problems. Furthermore, the SVM model performed the best when it was trained with all the verbalization and speech features together than with each feature separately. In addition, the SVM model's precision and recall

measures were also more balanced when it was trained with all the features together than with each feature separately.

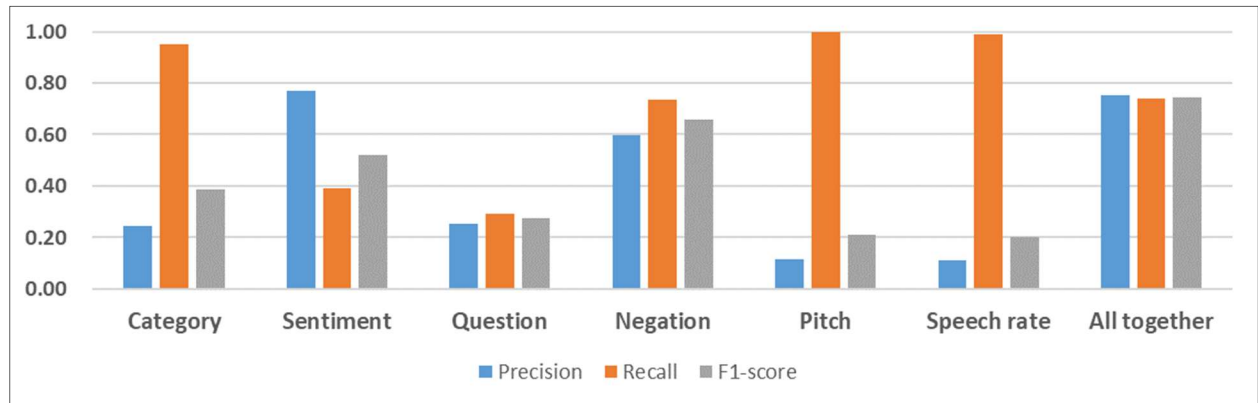


Figure 39. The precision, recall, and F1-score of the SVM models trained with each verbalization or speech feature respectively and together and evaluated using the 10-fold cross validation.

7.3.2 The Effect of *Products* on the Detection of Usability Problem Encounters

In practice, UX practitioners often work for a company to improve the user experience of a specific product at any given time. To reduce the workload of analyzing the think-aloud session data of a new test user, it is valuable to explore whether it is possible to build an ML model for *a product* using the data of the users who have interacted with the product to detect the usability problems encountered by *a new user* when the user uses the product?

To answer this question, I adopted the *leave-one-user-out* scheme to train and evaluate an SVM model for each *product*. I used SVM as it performed best among in terms of the F1-score and the balanced precision and recall from the cross-validations. For each of the four test products, I trained an SVM model using *the transcript (i.e., TF-IDF) + verbalization & speech features together* as input on any seven users' data and then tested the model using the rest one user's data. The rest one user was used to simulate the new user whose data the SVM model did not see in each round of evaluation. As each product was used by eight users, I repeated this evaluation

process eight times for each product so that each user was used as the new user once for the product. Finally, I averaged the measures of the SVM models across users for each product.

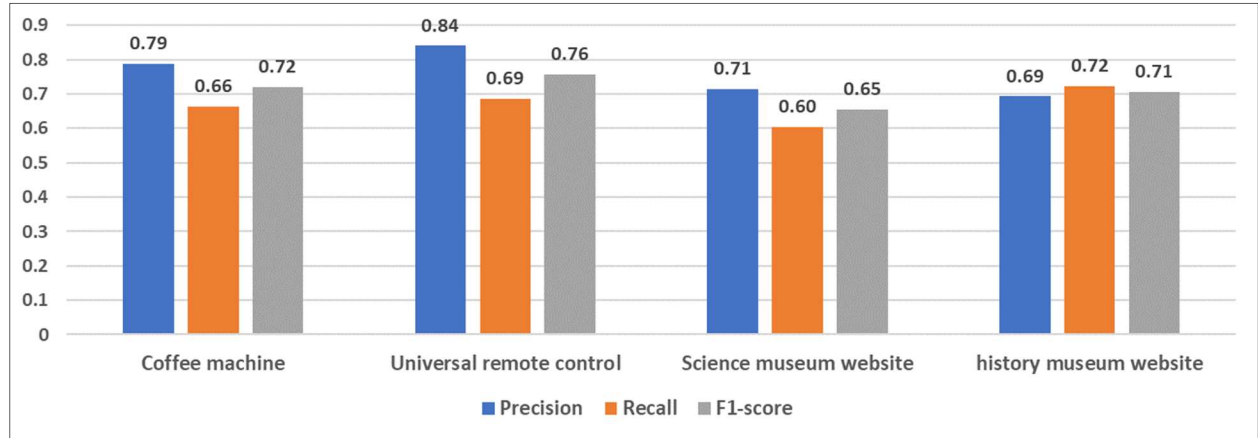


Figure 40. The average precision, recall, and F-1 score of the SVM model trained on any seven users' data using the *transcript* (i.e., *TF-IDF*) + *all the verbalization & speech features together* as the input and evaluated on the rest one user's data for each *product* respectively (i.e., *leave-one-user-out* scheme).

Figure 40 shows the average precision, recall, and F1-score of the SVM models for each product when using *the transcript feature* (i.e., *TF-IDF*) and *the verbalization and speech features together* as the input. It shows that it is possible to detect the usability problems for each product with reasonable precision, recall, and F1-score. In addition, the average F1-score of the SVM models for two physical devices and two digital websites were .74 and .68 respectively, which indicates that the models performed relatively better for the physical devices than for the digital websites.

7.3.3 The Effect of Users on the Detection of Usability Problem Encounters

It is common for companies to maintain a pool of participants whom they could contact over time for usability testing to reduce the recruitment cost. It is possible for companies to accumulate a dataset of the same user interacting with various products. If a company could build an ML model for a user using her thinking aloud data with existing products to predict when she would encounter

problems while interacting with a new product, the automatically detected encounters of the usability problems could help UX evaluators with their analysis. Therefore, with a dataset of a user interacting with different products, I was curious to understand whether it is possible to build an ML model for *a user* using the data of the products that the user has interacted to detect the usability problems encountered by the user when she uses *a new product*?

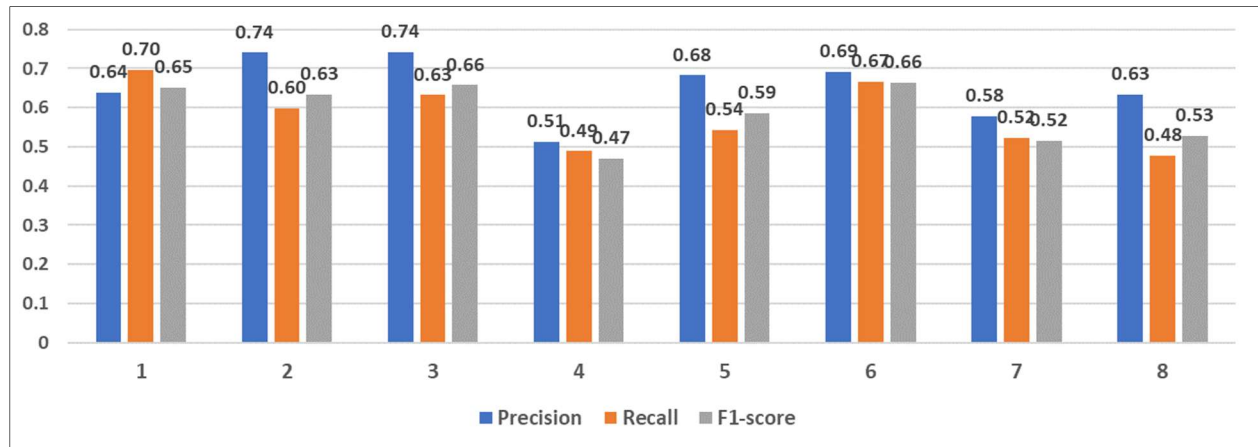


Figure 41. The average precision, recall, and F-1 score of the SVM model trained on any three products' data using *the transcript (i.e., TF-IDF) + all verbalization & speech features together* as input and evaluated on the rest one product's data for each *user* respectively (i.e., *leave-one-product-out* scheme).

To answer this question, I adopted the *leave-one-product-out* scheme to train and evaluate the ML model for each user respectively. I used SVM for this task again as it performed best among in terms of the F1-score and the balanced precision and recall. For each user in the dataset, I trained an SVM model for the user using *the transcript (i.e., TF-IDF) + verbalization & speech features together* as input on any three of the four products' data and tested the model on the rest one product's data. The rest one product was used to simulate the new product that the user has not yet used in each round of evaluation. As each user used four products in the dataset, I repeated this process four times so that each product was used as the new product once for each user. Finally, I averaged the measures across all the products for each user. Figure 41 shows the result. The F1-

score for the eight users ranged between .47 and .66. The performance was relatively better for some users (e.g., user 1, 2, 3, and 6) than others (e.g., user 4 and 7). This suggests that ML models can be built for individual users to detect their usability problem encounters. However, the effectiveness of these models is *user-dependent*.

7.3.4 Summary of the Key Findings

The following key points have emerged through the evaluations: 1) ML models trained using the *transcript feature* (i.e., TF-IDF or word embedding) can detect usability problem encounters; 2) Each *verbalization and speech feature* that tends to occur when users encounter usability problems, discovered via the studies described in Chapter 4, 5, and 6, can be used to improve the ML models' performance, and the best performance was achieved when *all* the verbalization and speech features were used together; 3) SVM performed the best in terms of F1-score and had the most balanced precision and recall among all four ML models. However, other ML models outperformed SVM in terms of precision or recall measure separately; 4) ML models can be built for each *product* using its existing users' data to detect the encounters of usability problems by a *new* user when she uses the same product, and 5) ML models can be built for each *user* using the data of the products that the user has interacted to detect the encounters of usability problems when using a *new* product.

7.4 Discussion

7.4.1 The *Verbalization and Speech Features* are Complementary in Detection of Usability Problem Encounters

The cross-validation results (Section 7.3.1) show that the *verbalization and speech features* that were shown to be indicative of usability problems in the studies (Chapter 4, Chapter 5, and Chapter 6) can also be used to improve the performance of the ML models for detecting the encounters of usability problems. The *verbalization and speech features* helped the most when they were used together than separately for training the ML models (Figure 37). It implies that the verbalization and speech features were complementary to each other. Furthermore, the results (Figure 38) also demonstrated that *the verbalization and speech features* can be used to train effective ML models

to detect the encounters of usability problems without needing to add the transcript feature (i.e., TF-IDF or word embedding). It implies that these verbalization and speech features were informative enough to capture the key characteristics of the encounters of the usability problems in the think-aloud sessions.

7.4.2 The *ML Models* for Detecting Usability Problem Encounters Have Different Precision and Recall Tradeoff

The cross-validation results (Section 7.3.1) also show that the SVM model performed best among all the four ML models in terms of F1-score, but it did not perform the best for either the precision or the recall. Furthermore, no single ML model performed the best for both precision and recall. Both RF and CNN outperformed SVM in terms of precision while falling short in their recall (.70) compared to SVM. This finding implies that instead of hoping to build a single ML model that performs the best in all measures (i.e., precision, recall, and F1-score), it would be better to develop different ML models and combine them to achieve the best performance.

Alternatively, UX practitioners should also consider whether precision or recall is more important in their particular context and choose the one that performs best for that measure. Indeed, a recent study that aimed to understand whether users would accept an imperfect Artificial intelligence (AI) [54] suggested that users valued precision and recall differently. This recommendation, however, would require UX practitioners to understand whether some ML models can perform consistently better than others in terms of precision, recall, F1-score, or other measures. On the other hand, it is also valuable to understand which measure (i.e., precision or recall) is more important in a particular context (e.g., a specific product or a user group) and choose the ML model that performs best for the more important measure.

Given the overwhelming evidence of the advantage of deep neural networks (DNNs) over the traditional ML methods, one might expect that CNN and RNN, would outperform to the SVM and RF models. The 10-fold cross-validation results, however, indicated that the opposite was often true. One potential reason might be that the dataset was relatively small which only contained eight users using four products and thinking aloud. In contrast, DNNs models are often data-hungry and

demand large amounts of data to learn the appropriate parameters to demonstrate their full potential. Future research should look into methods to effectively curate larger dataset that would allow the DNNs to learn their best parameters. However, curating larger dataset in the usability testing domain is challenging in practice because scheduling and conducting usability studies (e.g., think-aloud sessions) with participants in a controlled lab environment are often labor-intensive and time-consuming. One potential opportunity to gain large amounts of usability testing sessions is through remote usability testing, in which users can participate remotely in their convenient environment without the burden of scheduling and coming to the labs. Remote usability testing is promising also because it is cost-benefit effective (e.g., [12,15]). For example, Andreasen et al. showed that remote synchronous usability testing is virtually equivalent to the conventional lab-based controlled user studies [4]. Another challenge with curating larger dataset lies in *transcribing* and *annotating* the test sessions to consolidate the *ground-truth labels* for usability problems. In this work, researchers took the burden of completing these steps. Future research should develop better tools and methods that either facilitate UX professionals to label the data efficiently or automate or semi-automate the labeling process, for example, by designing appropriate crowd-sourcing approaches.

7.4.3 The Types of *Products* and *Tasks* Affect the Detection of Usability Problem Encounters

The leave-one-user-out evaluation for each product (Section 7.3.2) demonstrated that it is possible to build an ML model for a product to detect problems that a new user encountered when using the product. The implication is that companies could utilize recorded think-aloud test sessions that they have collected so far for a *product* to train an ML model to process the think-aloud sessions of a new user to pinpoint where in the sessions the new user encountered problems.

The leave-one-user-out evaluation results also show that the performance of the models was relatively better for the physical products (with the average F1-score of .74) than for the digital websites (with the average F1-score of .68). One potential reason for the difference might be the natural difference in the utterance of users when using physical and digital products. Another

possibility might be related to *the type of tasks* that users worked during the tests. For the physical products, users worked on *guided tasks* because they had access to the instruction manuals, which offer a prescribed set of steps to complete the tasks. In contrast, for digital websites, users worked on *guideless tasks* because they had no access to any prescribed steps. Although users can deviate from the prescribed steps when working on *guided tasks*, the availability of the guided steps could have influenced users' usage patterns and caused them to verbalize more similar utterances than when they were using digital products, for which they must freely explore the webpages to complete the *guideless tasks*. However, further research is needed to examine whether and how *the type of tasks* that users work on during think-aloud sessions influence their verbalizations and its implications on the design of effective ML models for detecting usability problems.

7.4.4 The *User-Dependent* Models for Detecting Usability Problem Encounters Vary in Performance

The leave-one-product-out evaluation for each user (Section 7.3.3) demonstrated that it is possible to build an ML model for each user using the data of the user interacting with existing products to detect problems that the user might encounter when using a new product. However, the result also shows a large variation in models' performance for different users. One potential reason for the variation in users' performance could be that different users may have verbalized their encounters of usability problems to different extents. Some users' verbalizations reflected the problems that they encountered more explicitly than other users. Another potential reason for the variation in the performance could be that some users may have verbalized their thought processes more consistently across products than other users. This consistency in their verbalizations may have helped the ML model learn and generalize. Further research is needed to determine the factors in users' verbalization and speech that lead to better performance for some users than others.

7.4.5 Automatic Verbalization Category Labelling

The goal of building and evaluating ML models was to understand the effect of the verbalization and speech features on the performance of detecting usability problem encounters. As a result, the verbalization category for each segment in the dataset was *manually* labeled to ensure the label's

accuracy. Since the evaluations show that the verbalization category is useful to improve the performance of the ML models, I took a further step to explore *whether it is possible to automatically detect the verbalization category for a segment based on its text content (i.e., the words that users uttered)*.

Informed by the finding of the three studies that the *Observation* category is more indicative of the usability problems among all the four categories, I sought to build a binary classifier to *detect whether a segment should be labeled as the Observation category or the non-Observation category*. To answer this question, I went through the category labels for all the verbalization segments and grouped the *Reading*, *Procedure*, and *Explanation* categories into the *non-Observation* category but kept the *Observation* label unchanged. I then followed the same procedure as described in sections 7.2.3 and 7.2.4 to compute the TF-IDF feature for each segment as the input feature and use the binary verbalization category labels as the ground truth to train an SVM classifier to classify whether a segment should be labeled as Observation or Non-Observation.

I performed 10-fold cross-validation on the SVM classifier using the entire dataset. The accuracy, precision, recall, and F1-score of the binary classifier were .83, .86, .71, and .78 respectively (*i.e.*, the first row in Table 19). The result implies that it is possible to reduce the effort of manually labeling the verbalization category, especially for large amounts of think-aloud sessions, by building a binary-class category classifier. Of course, to create such a classifier, UX practitioners still need to label a small portion of their data to curate the training data.

Although the binary verbalization classifier is enough for usability problem encounter detection, the other three verbalization categories (*i.e.*, *Reading*, *Procedure*, *Explanation*) could be useful in terms of providing contextual information to understand the issues that may ultimately be verbalized in an *Observation* segment, as the studies indicate. Therefore, it is also valuable to distinguish the four verbalization categories (*i.e.*, *Reading*, *Procedure*, *Observation*, and *Explanation*).

I further trained an SVM classifier to detect the four verbalization categories and performed a 10-fold cross evaluation on the entire dataset. The average accuracy, precision, recall, F1-score of the four-class classifier were .75, .78, .64, and .68 respectively (Table 19). Although the measures for four-category classification, as expected, are lower than those of the binary verbalization category classifier, the measures are nevertheless still very promising. Future work may explore more effective methods to improve the verbalization category detection accuracy, for example, by creating more effective features or ML models.

Table 19. The accuracy, precision, recall, and F1-score of the two-class (i.e., Observation and non-Observation) and four-class (i.e., Reading, Procedure, Observation, and Explanation) category classifiers.

	Accuracy	Precision	Recall	F1-score
Two-class classifier	.83	.86	.71	.78
Four-class classifier	.75	.78	.64	.68

7.5 Summary

Fast-pace analysis for recorded think-aloud sessions is needed to alleviate the burden of usability evaluators. Toward this goal, I took the first step to design and evaluate methods to automate the detection of usability problem encounters in think-aloud test sessions. The evaluations show that when using *the verbalization and speech features* (i.e., category, sentiment, question, negation, abnormal pitch, and abnormal speech rate) that are shown to be linked to the usability problems as the input, the performance of four different ML models improved compared to only using the basic transcript feature (i.e., TF-IDF or word embedding) as the input.

Furthermore, the evaluations show that it is possible to build an ML model for a *product* using its existing users' data to detect the usability problems encountered by a new user; it is also possible

to build a user-dependent ML model for a *user* to detect usability problems encountered by the user when she uses a new product.

Last, although recent research shows that UX practitioners often struggle to understand the capabilities and limitations of ML [25,104], some also suggest that UX practitioners can design ML-enhanced products without knowing thoroughly about the ML models [103]. This is encouraging because the ultimate goal of building ML models to detect usability problem encounters is to help UX practitioners identify usability problems more effectively. Therefore, it is important to explore how to best present the ML models or their detection results to UX practitioners so that they can leverage the power of the ML without being overwhelmed in making their analysis.

In the next chapter, I will describe how I took the first step to explore ways to present the M-inferred usability problem encounters to UX practitioners and understand whether and how the UX practitioners perceive, interact, and incorporate the ML-inferred usability problem encounters into their analysis.

Chapter 8 Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions

I have demonstrated that computational models can be built to predict the usability problem encounters with promising accuracy. It is important to explore further how to make use of the computationally predicted usability problem encounters to help usability evaluators identify problems more effectively. In this chapter, I explore the last research question of this dissertation:

RQ4: Can UX practitioners use ML inferences of usability problem encounters to help them with their analysis?

To answer RQ4, I designed and evaluated an intelligent visual analysis tool that visualizes the ML-inferred usability problem encounters as well as the ML's input features. Through a controlled user study, I demonstrate that UX practitioners can identify more usability problems and pay more attention to the areas of the think-aloud sessions where they would otherwise. This is achieved by showing the timeline of ML-inferred usability problem encounters, which the practitioners would use as overviews, reminders, and anchors, and the timeline of ML input features, which the practitioners would use to better understand the features that the ML considers and the potential mistakes that the ML might make and to better allocate their attention.

8.1 Research Questions

I explore the following four sub-questions, which tackle different aspects of the interaction between UX evaluators and the ML-inferred usability problem encounters. The findings of these sub-questions together answer the RQ4:

- RQ4-1: Would ML-inferred usability problem encounters improve UX practitioners' efficiency? How would UX practitioners leverage ML in their analysis?
- RQ4-2: How would ML-inferred usability problem encounters influence UX practitioners' analysis strategies? Would they tend to review video sessions with more rewinds or pauses?

- RQ4-3: How would UX practitioners perceive and manage the relationship with the ML? How would they deal with disagreements and limitations of the ML-inferred problem encounters?

To answer these three sub-questions, I first iteratively developed a visual analytics tool—VisTA—that integrates machine intelligence and then used VisTA as a vehicle to answer the three sub-questions. I designed a controlled lab study to expose UX practitioners to ML at different levels and recorded a rich set of quantitative and qualitative data about their interactions with the ML-inferred problem encounters, their analysis behaviors (e.g., pauses and rewinds), and their perceived relationship with ML.

In the rest of the chapter, I first describe the dataset and the ML models for detecting usability problems. I then describe how I designed the visual analytics tool. Next, I introduce the study design and analysis methods. Finally, I present the study results and discuss the implications of the findings.

8.2 Think-Aloud Dataset and Problem Encounter Detection

8.2.1 Dataset

Three think-aloud sessions from Study 3 (i.e., the generalization study described in Chapter 6) were randomly chosen for each evaluator to analyze. To ensure that there would be no learning effect between sessions, these three think-aloud sessions were about three different participants using three different products. Specifically, in these three recorded think-aloud sessions, each of the three participants worked on a task on one of the three products, one digital product (i.e., the national science and technology museum website) and two physical products (i.e., one universal remote control and one multi-function coffee machine).

All think-aloud sessions were video recorded with the audio stream. The average session duration was 222 seconds ($\sigma = 131$) for the website, 619 seconds ($\sigma = 195$) for the universal remote control, and 854 seconds ($\sigma = 251$) for the coffee machine.

8.2.2 Data Labelling and Feature Extraction

The details of how the data and the features were labeled or computed have been described in Study 3 (i.e., the generalization study described in Chapter 6). For completeness of the chapter, I describe the core steps in data labeling and feature extraction here.

The think-aloud sessions were manually transcribed into text. Then, two coders followed a similar approach used in previous work [24,30] to divide each think-aloud session recording into small segments. The beginning and end of a segment were determined by pauses between verbalizations and the verbalization content [24,30]. Each segment corresponded to a verbalization unit, which could include single words, but also clauses, phrases and sentences. For each segment, two coders first labeled independently whether the think-aloud user experienced a problem (e.g., being frustrated, confused or experiencing a difficulty) and later discussed to consolidate their labels. I used the binary problem labels as the *ground truth* for training ML models.

For each segment, two coders assigned it with one of the four verbalization categories (i.e., reading, procedure, observation, and explanation) [24]. The results of the research in Chapter 4, Chapter 5, and Chapter 6 have shown that when users experience problems in think-aloud sessions, their verbalizations tend to include the *Observation* category, *negative sentiment*, *negations*, *questions*, *abnormal pitches*, and *speech rates*. Inspired by this finding, in addition to labeling the *category* information for each segment, I computed its *sentiment* based on the transcript using the VADER library [48]. Moreover, I designed a keyword matching algorithm to determine whether users verbalized *negations* (e.g., no, not, never) in a segment. Similarly, I designed a keyword matching algorithm to determine whether users *asked a question* in a segment by searching for keywords (e.g., what, when, where) that were located at the beginning of a sentence. Lastly, for each segment, I computed user's *pitch* (HZ) using the speech process toolkit Praat [9] and the *speech rate* by dividing the number of words spoken in a segment by its duration. To determine whether the user verbalized with abnormally high or low pitches and speech rates, I computed the mean and the standard deviation (STD) of the pitch and the speech rate of the entire think-aloud session and

automatically labeled a segment as having *abnormally high or low pitch or speech rate* if any value in the segment was two standard deviations higher or lower than the mean pitch or speech rate.

In sum, six *verbalization* features were generated for each segment: category, sentiment, negations, questions, abnormal pitches, and abnormal speech rates. In addition, for each segment, I also computed the *TF-IDF* (i.e., term frequency-inverse document frequency) using scikit-learn library [82] and trained *word embeddings* on the dataset using Tensorflow [1]. In the end, eight features were used as the input for training a range of machine learning models to determine whether the user encountered a problem in each segment.

8.2.3 Model Training and Evaluation

I employed four machine learning methods: random forest (RF), support vector machine (SVM), convolutional neural network (CNN), and recurrent neural network (RNN), which have been shown effective in text-based classification tasks.

I extracted the TF-IDF features from the data set and used them to train SVM and RF models using the scikit-learn. I used the word-embedding features to train CNN and RNN models. I used ReLU as the activation function for CNN. For RNN, I used Gated Recurrent Units (GRU) as cells and used SoftMax as the activation function. To evaluate the models, I performed a 10-fold cross validation on the data set and used the performance of these models as the *baseline*.

In addition, I combined the TF-IDF and word embedding features (i.e., TF-IDF or word-embedding) with the six verbalization features (i.e., category, sentiment, negations, questions, abnormal pitches, and abnormal speech rates) as the input to train the same four ML models. Similarly, I performed 10-fold cross-validation on the dataset.

As shown in Table 20, the results indicate that verbalization features helped improve the performance for all ML models and the SVM models performed the best on the dataset. Thus, I decided to train SVM models using the TF-IDF or word embedding with the six verbalization features and use them to predict the usability problem labels (i.e., whether the user experienced a

problem or not) for all segments of each think-aloud session. After this process, all segments in each think-aloud session had an ML-inferred problem label. These ML inferred problem labels and the features were then used in the visual analytics tool, which will be described in the next section.

Table 20. The performance of the ML models when trained with different sets of features.

	TF-IDF/Word embedding			All features		
	Precision	Recall	F-score	Precision	Recall	F-score
RF	0.79	0.53	0.64	0.80	0.64	0.71
SVM	0.59	0.73	0.65	0.76	0.70	0.73
CNN	0.81	0.41	0.54	0.79	0.48	0.60
RNN	0.60	0.43	0.50	0.76	0.54	0.64

8.3 VisTA: Visual Analytics Tool for Think-Aloud

Following a typical user-centered iterative design process, I developed VisTA to interactively present the verbalization features and the usability problems detected by ML as described earlier.

8.3.1 Design Principles

In the initial study (Chapter 4), the confirmation study (Chapter 5), and the generalizations study (Chapter 6), usability evaluators had access to a visualization of the verbalization and speech features of the entire think-aloud session (see Figure 18, Figure 19, Figure 30, and Figure 31) as a series of synchronized timelines in addition to the functions that allow evaluators to play the recorded sessions and add problem descriptions. Figure 42 shows a close-up snapshot of the feature panels.

Based on the feedback received from usability evaluators on their experiences and preferences for the interface, I derived two principles to improve the visual presentation of the verbalization and speech features.

- *Be Simple and Informative*

Evaluators wanted to have a simple interface that offers concise information that they could consult to if need, while allowing them to focus on watching or listening to the recorded sessions. Although evaluators felt that each of the feature can be informative, showing all of them at once was overwhelming, as one evaluator pointed out that “*because lines are so busy, it is hard to pick up significant areas while reviewing the session.*” Instead of viewing all the raw features and trying to figure out important information, they would prefer just having one *condensed* type of information while still being able to access the raw features if needed.

- *Be Interactive and Responsive*

Evaluators felt that the function of clicking anywhere on any timeline to move the session recording to that timestamp was helpful. In addition, they wanted to interact with the input features, such as filtering particular features, to better understand and leverage the features. Evaluators also wanted to tag their identified problems with short annotations to facilitate their analysis.



Figure 42. The initial way of visualizing verbalization and speech features.

I adopted these two principles in the design of VisTA. Specifically, I integrated the machine intelligence into the analysis flow among other capabilities (Figure 43). The refined VisTA interface provides a typical video player, a *problem timeline* that visualizes the ML's inferred

problems, and a *feature timeline* that visualizes ML's input features on the left side, as well as a panel on the right side for logging and tagging identified problems and filtering input features.

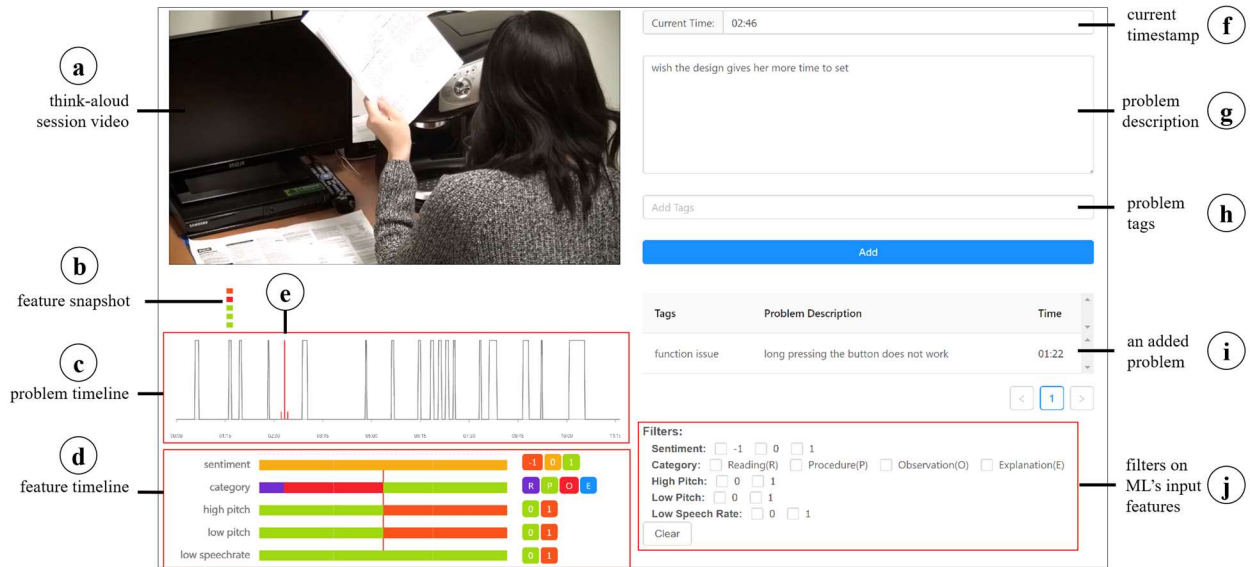


Figure 43. VisTA: a visual analytics tool that allows UX practitioners to analyze recorded think-aloud sessions with the help of machine intelligence to detect usability problems.

8.3.2 Session Reviewing and Problem Logging

Usability evaluators can play and pause the think-aloud video (Figure 43a) by pressing the ESC key or fast-forward or backward by pressing the right or left arrow keys on the keyboard. While the video is playing, the current timestamp in Figure 43f automatically updates. Evaluators can write a problem description in Figure 43g, add tags to the problem description in Figure 43h, and finally log the problem by pressing the “Add” button.

All problems identified so far are visualized in the table in Figure 43i. Clicking an added problem entry in the table navigates the video to the timestamp on the timeline where the problem was added so that evaluators can reply to the problem segment video if needed.

Moreover, the tag area in Figure 43h allows evaluators to create multiple tags and attach them to a problem description. VisTA records the tags that evaluators have created so far and shows the tags to the evaluators in a dropdown list for reuse.

8.3.3 Visualization of Problem Encounters and Features

VisTA visualizes ML-inferred problem encounters on the *problem timeline* (Figure 43c), following the design principle “be simple and informative” and the idea of showing “condensed” information (see 8.3.1). The design of a timeline to show only the predicted problem encounters hides the complexity of the raw verbalization and speech features that are hard and overwhelming for evaluators to understand in their analysis. Because this is the primary augmented information to a think-aloud session video, it is placed directly under the video player to facilitate quick scanning. The long red vertical line on the problem timeline as shown in Figure 43e indicates the current time of the video.

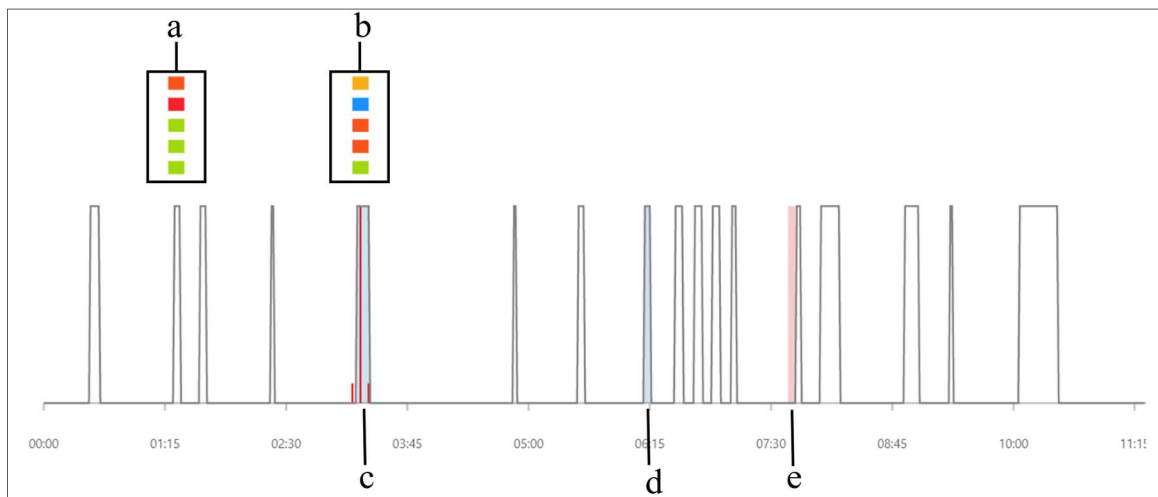


Figure 44. The *problem timeline* of VisTA. The problem timeline highlights all the segments (i.e., c, d, and e in the chart) that have the same features as the currently paused timestamp (c), which allows evaluators to examine where in the session the same features appeared and how those areas align with the ML-inferred problems.

Each ML-inferred problem encounters are visualized as a “spike” on the problem timeline (Figure 44). Some spikes are wider than others. This is because the ML predicts whether the user encounters a problem or not for each segment independently and the segments can have varying lengths in time. As was described in the second paragraph in section 8.2.2, the length of a segment was determined by the pauses between segments and the actual content verbalized in the segment. For example, it is likely that there were many shorter segments between 06:15 and 07:30 on the problem timeline and one longer segment right after 10:00. Another reason for having a wider spike on the problem timeline is that the duration of the spike might contain many small consecutive segments that were all labeled by the ML as having problems.

Further, to allow evaluators to access the raw features without being overwhelmed, VisTA only reveals the main input features in a short time window (i.e., five seconds before and five seconds after the current time) around the current time in the video, instead of the entire video as in the initial designs used in the three studies (i.e., Chapter 4, Chapter 5, and Chapter 6), on the *feature timeline* as shown in Figure 43d and Figure 45. The start and end of the window are marked with two short red vertical lines on the problem timeline with the current time marked as the long red vertical line (Figure 43e and Figure 44c). When evaluators play the video, the feature values on the feature timeline are dynamically updated. As this is a less demanded feature per evaluators' feedback, it is placed under the problem timeline.

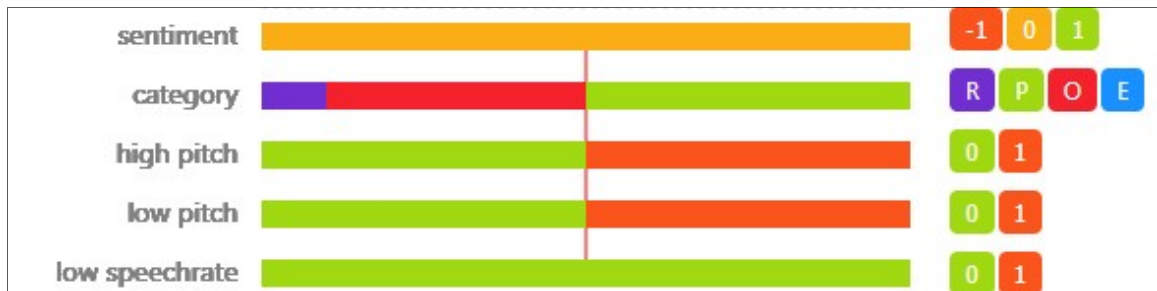


Figure 45. The feature timeline that shows the ML’s input features in a short time window around the current time in the video. The current time in the video is marked as the red vertical line in the center.

When an evaluator pauses the video, VisTA shows a snapshot of the input features at the current time on the top of the problem timeline as a stack of colored bars (Figure 44b). It also highlights parts of the video that have the same features. For example, Figure 44 c, d, and e contain the same features as the current time as shown in Figure 44b. To help evaluators better assess how these highlight areas align with the ML-inferred problem encounters, the highlight areas where ML also detects problems are color-coded in blue (Figure 44c,d) to hint that the ML also thinks that there is a problem and those where ML detects no problems are color-coded in pink to hint that the ML does not think there is a problem (Figure 44e). When the video is playing again, the highlight and the feature snapshot will disappear to avoid potential distraction.

VisTA adds a feature snapshot on the top of the problem timeline (Figure 44a) at the time when evaluators add a problem to help them remember the locations of the problems that they have added so far and what the features for each problem look like. When evaluators click on the snapshot, VisTA highlights all areas that have the same set of features on the problem timeline.

Furthermore, as shown in Figure 43j, VisTA also provides a filter function that allows evaluators to manually select a combination of features, which automatically highlights the areas on the problem timeline that have the same set of features. I hypothesize that the highlighting areas would allow evaluators to better assess how the features of their choice align with the ML-inferred problems.

8.3.4 Implementation Details

The application was implemented in JavaScript using the React.js library. I used the Ant design library to render basic UI components (e.g., textbox, button, and table), used the BizChart library to visualize the ML-inferred problem chart and the ML's input features chart, and used the video-react.js library for the video player. I used the MobX.js library to record inputs and events on the UI and used Node.js for backend services, such as transmitting videos.

8.4 User Study

8.4.1 Study Design

To investigate how UX practitioners would use the problem timeline and the feature timeline in their analysis, I conducted a controlled laboratory study to compare different versions of VisTA.

More specifically, I developed *VisTASimple* that only shows the *problem timeline without the input features* (Figure 46), in order to better understand the effect of the feature timeline on a UX practitioner's analysis. This also allows for investigating how it can affect the user interactions on the problem timeline.

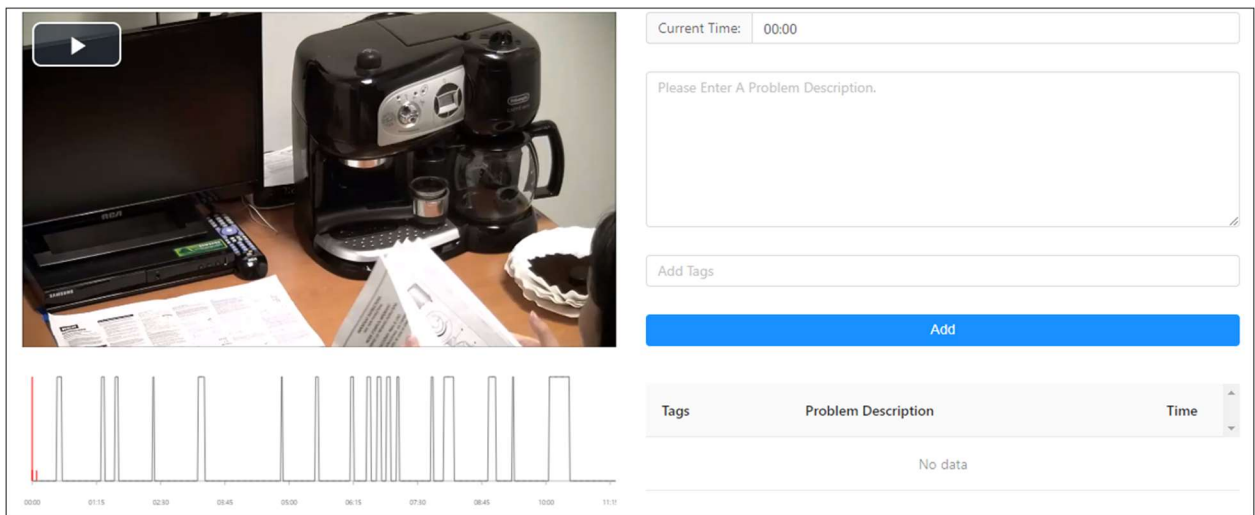


Figure 46. VisTASimple UI. Compared to VisTA, VisTASimple only presents the problem timeline without showing the ML’s input features or providing the feature filtering function.

Moreover, to study the effect of the whole ML in the analysis process, I included a *Baseline* condition that shares the same user interface as VisTASimple except *not having the problem timeline*. Figure 47 shows the Baseline interface. When using the Baseline interface, UX practitioners do not have any access to ML.

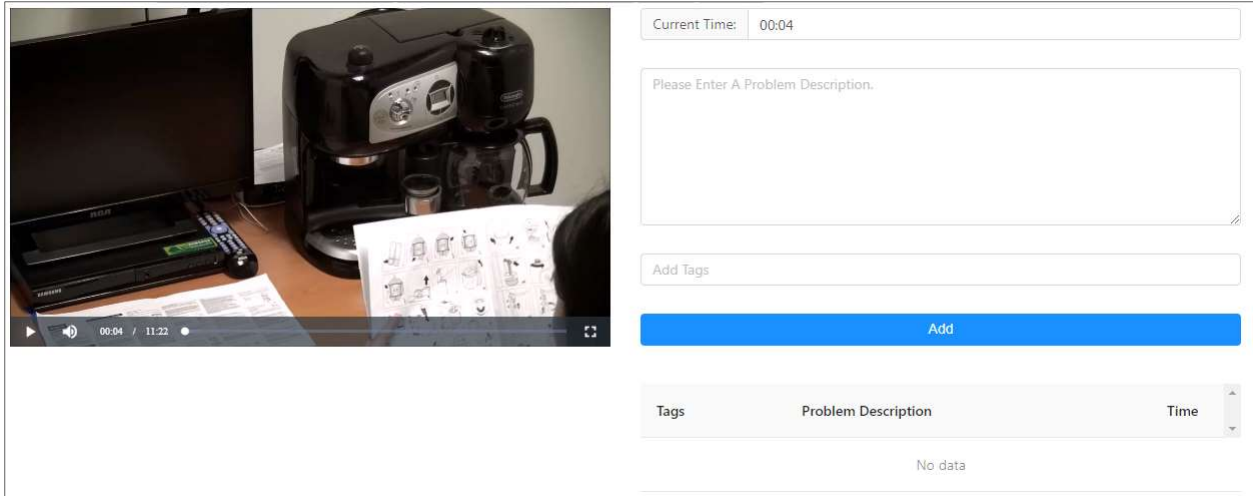


Figure 47. Baseline UI. Baseline has the same video reviewing and problem annotation functions as the VisTA and VisTASimple but it has no ML-related features.

Because there are potentially learning effects between conditions, I adopted a between-subjects design for the study. For example, after a participant used VisTA, she would know the input features of the ML, which might prime her to consider these features in the other two conditions.

8.4.2 Participants

I recruited 30 UX practitioners from local UX communities at a large metropolitan area by posting advertisements on social media platforms. They participated in the study as *usability evaluators*. I randomly assigned them to the three conditions, thus each having 10 evaluators. They reported their years of experience as a UX practitioner ranging from one to nine years. The averages for the Baseline, VisTASimple, and VisTA conditions were the same: 3 years, with standard deviations of 2, 3, and 2 years respectively. Mann Whitney U test found no significant difference in the years of experience among the three conditions.

8.4.3 Procedure

I conducted the studies as the study moderator in a quiet office room with a 27-inch monitor connected to a laptop computer. After getting the evaluators' informed consent, the moderator

explained that their task was to review three recorded think-aloud sessions to identify when users were confused, frustrated, or experienced problems. The three videos were about users operating on three different products (*i.e.*, one website, one universal remote, and one coffee machine), and were randomly chosen from the dataset described in Section 8.2, and the same set of videos was used for all participants in the whole study.

At the beginning of the study, evaluators were demonstrated how to use the tool (*i.e.*, Baseline, VisTASimple, or VisTA) by loading a trial think-aloud session, and the moderator answered any questions that they had.

In each session, before evaluators analyzing the video, the moderator introduced the product and the task that the user worked on in the recorded video. Evaluators had to finish the analysis within a maximum of three times of the video playback length. After each session, the moderator conducted a brief interview by asking how they analyzed the video.

At the end of the whole study, evaluators filled in a questionnaire to rate their experience in using ML (for VisTA and VisTASimple) and their confidence in the problems that they identified on a 7-point Likert scale. Then, the moderator interviewed evaluators to further understand their confidence in the analysis results and their usages to the problem and the feature timelines (where appropriate).

All interviews were audio-recorded. The study lasted about 1.5 hours, and each evaluator was compensated with \$30.

8.5 Analysis and Results

I describe how the data were analyzed and present the key findings in this section.

8.5.1 Data Capture and Analysis

The software tool in all three conditions (*i.e.*, Baseline, VisTASimple, and VisTA) recorded evaluators' interactions during the study. Specifically, it saved all the problem descriptions,

problem tags, and their corresponding timestamps. I analyzed these data to understand the problems that evaluators identified. The tool also continuously recorded pairs of timestamps per second, (*SessionTime*, *VideoTime*), when evaluators were analyzing the sessions. This reflects the relationship between the timestamps in the video and in the study session. I analyzed this information to understand how evaluators interacted with the videos to identify problems. The results of the quantitative analysis will be reported in Section 8.5.3

In addition, all interviews with the evaluators were recorded and transcribed. Two researchers coded the transcripts independently and then discussed to consolidate their codes. They then performed affinity diagramming to group the codes and identify the core themes emerged from the data. The findings from analyzing these qualitative data will be described in Section 8.5.4.

8.5.2 Overview of the Quantitative and Qualitative Results

I provide quantitative and qualitative data in the next two sections to answer the three sub research questions of RQ4, which were introduced in the beginning of this chapter. I reiterate the three sub research questions as follows:

- RQ4-1: Would ML-inferred usability problem encounters improve UX practitioners' efficiency? How would UX practitioners leverage ML in their analysis?
- RQ4-2: How would ML-inferred usability problem encounters influence UX practitioners' analysis strategies? Would they tend to review video sessions with more rewinds or pauses?
- RQ4-3: How would UX practitioners perceive and manage the relationship with the ML? How would they deal with agreement, disagreements, and limitations of the ML-inferred problem encounters?

Specifically, Section 8.5.3.1 (Problem Identification), Section 8.5.4.1 (How did evaluators use the problem timeline?), and Section 8.5.4.2 (How did evaluators use the feature timeline?) answer RQ4-1; Section 8.5.3.2 (Session Review Strategies) answers RQ4-2; and Section 8.5.3.3 (Questionnaire), Section 8.5.4.3 (What were evaluators' attitudes toward ML-inferred usability

problem encounters?), Section 8.5.4.4 (How did evaluators deal with agreement and disagreement with ML-inferred usability problem encounters?), and Section 8.5.4.5 (What did evaluators perceive as the limitations of ML-inferred usability problem encounters?) answer RQ4-3.

8.5.3 Quantitative Results

8.5.3.1 Problem Identification

To answer RQ4-1, I counted the number of problems identified in each condition for each session (Table 21). Evaluators found the highest number of problems when using VisTA, followed by VisTASimple and then Baseline. VisTA presents the ML-inferred usability problem encounters as well as the ML's input features in addition to the basic video review and annotation functions. VisTASimple presents the ML-inferred usability problem encounters only in addition to the same set of video review and annotation functions. Baseline presents no ML related information but only the same set of video review and annotation functions. One-way ANOVA found no significance in the number of problems identified between conditions for the first ($F(2,27)=2.70$, $p=.09$, $\eta_p^2=.17$), and the second session ($F(2,27)=1.33$, $p=.28$, $\eta_p^2=.09$), but found a significant difference for the last session ($F(2,27)=4.13$, $p=.03$, $\eta_p^2=.23$). Post-hoc Bonferroni-Dunn test found a significant difference between Baseline and VisTA.

Table 21. The number of problems identified by evaluators ($\mu(\sigma)$)

	Session 1	Session 2	Session 3
Baseline	3.9 (2.0)	6.2 (2.7)	13.8 (4.8)
VisTASimple	5.9 (2.8)	7.1 (2.4)	18.2 (6.6)
VisTA	6.7 (3.3)	8.4 (3.8)	21.2 (5.9)

Two researchers went through each problem that evaluators added and compared the descriptions and timestamps with the ML-inferred problems. The agreement and disagreement of the identified problems between evaluators and ML are indicated in Table 22.

One-way ANOVA found no significant difference in the number of problems that evaluators and ML agreed for the first session ($F(2,27)=1.6$, $p=.22$, $\eta_p^2=.1$) and the second session ($F(2,27)=.9$,

$p=.42$, $\eta_p^2=.06$), but found a significant difference for the third session ($F(2,27)=5.8$, $p=.008$, $\eta_p^2=.30$). Post-hoc Bonferroni-Dunn test found a significant difference between Baseline and VisTA. In contrast, there were no significant difference in the number of problems that were identified only by evaluators for the first ($F(2,27)=.07$, $p=.93$, $\eta_p^2=.005$), the second ($F(2,27)=.006$, $p=.99$, $\eta_p^2=.0005$), or the last session ($F(2,27)=2.2$, $p=.13$, $\eta_p^2=.14$). Similarly, there were no significant difference in the number of problems that were identified only by ML for the first ($F(2,27)=2.3$, $p=.12$, $\eta_p^2=.15$), the second ($F(2,27)=.66$, $p=.52$, $\eta_p^2=.05$), or the last session ($F(2,27)=1.6$, $p=.22$, $\eta_p^2=.11$).

Table 22. The agreement and disagreement of identified problems between evaluators and ML. 😊🤖: problems that evaluators and ML agreed; 😊: problems that only evaluators identified; 🤖: problems that only ML identified. Results are shown as $(\mu(\sigma))$.

	Session 1			Session 2			Session 3		
	😊🤖	😊	🤖	😊🤖	😊	🤖	😊🤖	😊	🤖
Baseline	2.0 (1.1)	2.2 (2.1)	3.8 (1.2)	3.4 (0.9)	3.4 (2.3)	0.9 (0.8)	9.3 (2.7)	4.4 (2.5)	8.4 (2.3)
VisTASimple	3.7 (1.7)	1.9 (1.5)	2.7 (1.5)	3.8 (2.0)	3.3 (2.1)	0.8 (0.8)	12.6 (4.2)	5.6 (3.8)	6.1 (3.1)
VisTA	4.0 (3.0)	2.1 (1.1)	2.3 (2.2)	5.1 (2.4)	3.3 (2.0)	0.4 (0.8)	15.1 (4.7)	6.8 (3.5)	4.0 (3.8)

8.5.3.2 Session Review Strategies

To answer RQ4-2, I analyzed how evaluators reviewed the recorded think-aloud sessions. Specifically, I counted the number of times that evaluators paused and rewind the video in each session under each study condition (Table 23). Results show that evaluators paused the most when using VisTA, followed by VisTASimple and then Baseline. One-way ANOVA showed that the difference was not significant for the first sessions ($F(2,27)=1.9$, $p=.17$, $\eta_p^2=.12$), but was significant for the second ($F(2,27)=4.6$, $p=.02$, $\eta_p^2=0.25$) and the third session ($F(2,27)=6.4$, $p=.006$, $\eta_p^2=.32$).

Table 23. The number of times for pauses and rewinds ($\mu(\sigma)$).

	Conditions	Session 1	Session 2	Session 3
Pauses	Baseline	5.0 (3.3)	5.4 (3.5)	7.6 (6.9)
	VisTASimple	8.4 (5.5)	8.0 (4.7)	17.1 (6.4)
	VisTA	8.7 (7.3)	12.3 (7.6)	17.5 (7.5)
Rewinds	Baseline	4.5 (3.8)	4.3 (3.2)	7.6 (5.8)
	VisTASimple	20 (15.3)	15.3 (12.3)	18.0 (9.4)
	VisTA	5.2 (6.2)	9.0 (6.7)	9.8 (6.8)

Further, evaluators rewind the most when using VisTASimple, followed by VisTA and then Baseline. One-way ANOVA indicated that the difference was significant for the first ($F(2,27)=7.7$, $p=.002$, $\eta_p^2=.36$), the second ($F(2,27)=3.4$, $p=.049$, $\eta_p^2=0.20$), and the third session ($F(2,27)=5.0$, $p=.014$, $\eta_p^2=.27$). The differences between VisTASimple and Baseline for all the three sessions were significant, but the differences in all other condition pairs were not significant. These results along with the results about the number of reported problems together show that evaluators were able to identify significantly more problems without needing to rewind the session videos significantly more often when using VisTA than Baseline.

To further understand evaluators' session reviewing behaviour, I analyzed the pairs of timestamps (*SessionTime*, *VideoTime*). I categorized typical evaluator behaviour by both *the number of passes on a video* and *the playback behaviour when going through a single pass* (Figure 48). In general, evaluators adopted one of the *one-pass* and *two-pass* approaches.

For the *one-pass* approach, there were three typical behaviours, namely *No-Pause-Write*, *Pause-Write*, and *Micro-Playback-Write*. *No-Pause-Write* means that evaluators kept the video playing while entering the problems identified (Figure 48a). This behaviour was more common in the third video potentially due to the video length and the number of problems presented. For *Pause-Write*, evaluators paused the video while they enter the problems identified (Figure 48b). With *Micro-Playback-Write*, evaluators repeatedly rewind and played a small section of the video while entering the problems identified (Figure 48c). Evaluators who used VisTA or VisTASimple tended to adopt the *Micro-Playback-Write* strategy more than the Baseline. In particular, this strategy was adopted 6 times in Baseline, 18 times in VisTASimple, and 11 times in VisTA across all the

sessions. It suggests that seeing the problem timeline had made them more cautious in their analysis. In addition, the Micro-Playback-Write strategy was adopted more in VisTASimple than VisTA, suggesting that knowing the input features of ML allowed them to trust ML more and thus needed to rewind less frequently.

When evaluators adopted the *two-pass* approach, some used the first pass to gain an understanding of the context and to get a heads up of where the problems might be, i.e., *Overview-then-Write*. They sometimes played through the video without pausing or rewinding in the first pass if gaining context was the goal (Figure 48d). On the other hand, some evaluators identified problems during the first pass and used the second pass as a chance to pick up the problems they might have missed or re-assessed issues they were not sure of, i.e., *Write-then-Check* (Figure 48e).

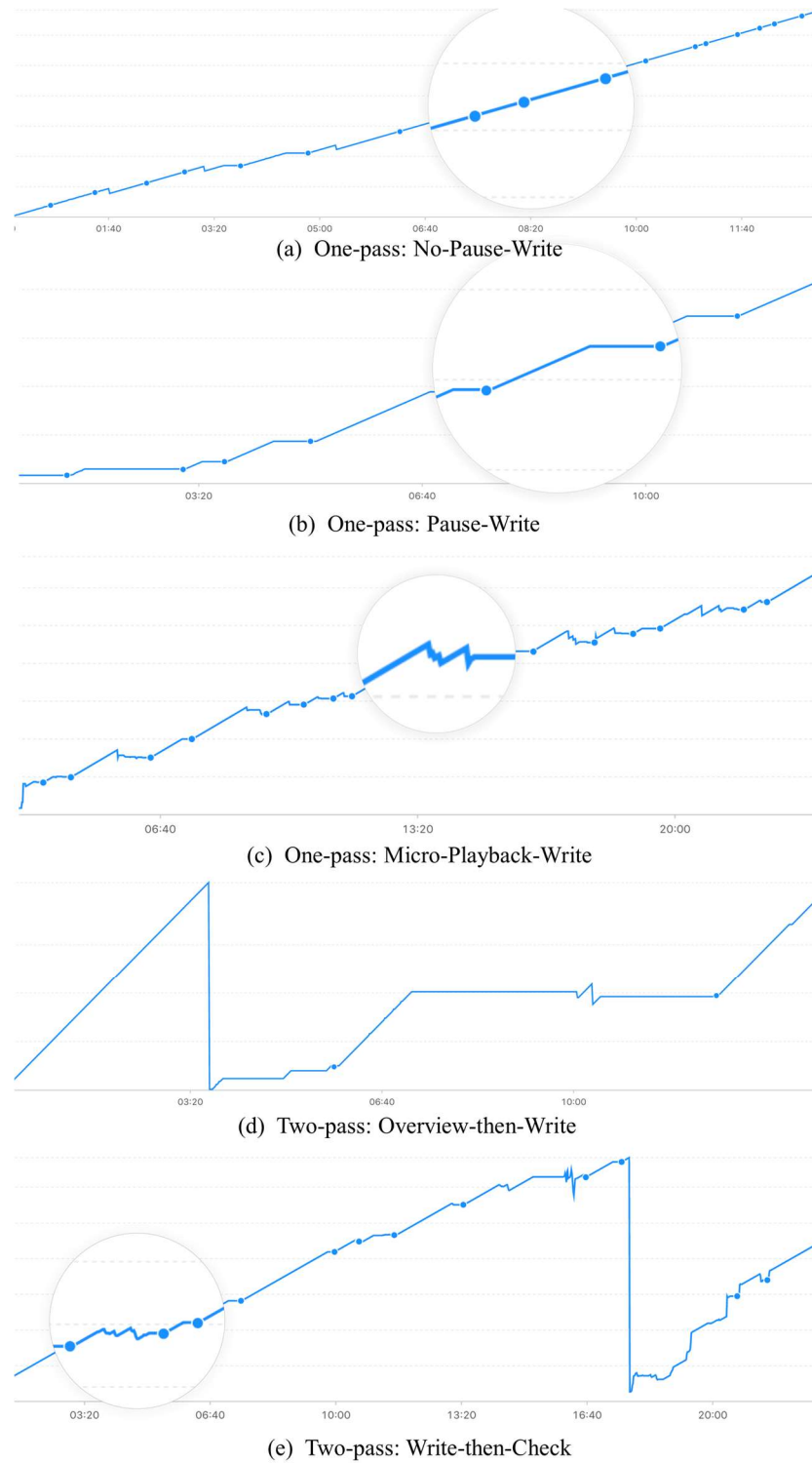


Figure 48. Typical playback behaviours (x-axis: session time; y-axis: reviewed video time).

Table 24 shows the number of evaluators who used the one-pass or two-pass approach. In all conditions, evaluators adopted one-pass, and two-pass approaches and the proportions of the two were similar between conditions. The two-pass behaviour was more common in the first session than the last two, which may be due to the length of the videos, as the last video was the longest among all.

Table 24. The frequency of evaluators' session reviewing strategies based on passes.

Conditions	Session 1		Session 2		Session 3	
	1-pass	2-pass	1-pass	2-pass	1-pass	2-pass
Baseline	6	4	8	2	10	0
VisTASimple	4	6	6	4	8	2
VisTA	7	3	9	1	9	1

8.5.3.3 Questionnaire

To answer RQ3, I analyzed the questionnaire responses regarding the usage of the tool and their confidence in identified problems. Evaluators strongly agreed that they compared the ML-inferred problems in their analysis when using VisTASimple ($Mo=7$, $Md=7$) and VisTA ($Mo=7$, $Md=6.5$). They felt positive that they knew how to make use of the problem timeline when using VisTASimple ($Mo=5$, $Md=6$) and VisTA ($Mo=6$, $Md=6$). In general, evaluators agreed that ML helped them notice parts of the videos that they might have skipped if analyzing the videos without it when using VisTASimple ($Mo=5$, $Md=5$) and VisTA ($Mo=5$, $Md=5$). Also, evaluators would like to use VisTASimple ($Mo=6$, $Md=6$) and VisTA ($Mo=5$, $Md=6$) in future analysis.

Interestingly, based on the ratings, I found that evaluators agreed more on the problems that ML identified ($Mo=5$, $Md=5$) than the problem-free areas that ML suggested ($Mo=3$, $Md=3$) when using VisTA, and the difference was significance ($z'=-1.98$, $p'=.047$). In contrast, there was no significant difference in the ratings of the two questions when using VisTASimple.

I found that evaluators were confident that others would agree on the problems they identified: Baseline ($Mo=6$, $Md=5.5$), VisTASimple ($Mo=6$, $Md=6$), and VisTA ($Mo=5$, $Md=5$). Kruskal-Wallis test found no significant difference ($H'=1.79$, $p'=0.40$). They were also confident about the

areas that they identified as problem-free: Baseline ($Mo=4$, $Md=5$), VisTASimple ($Mo=5$, $Md=5$), and VisTA ($Mo=6$, $Md=6$). No significant difference was found between conditions ($H'=2.85$, $p'=0.24$).

8.5.4 Qualitative Results

To further answer RQ4-1 and better understand how evaluators were able to identify more problems with VisTA, which presents both ML-inferred usability problem encounters on the problem timeline and the ML's input features on the feature timeline, I provide insights from the interview data about evaluators' interactions with the problem timeline (section 8.5.4.1) and the feature timeline (section 8.5.4.2). To better answer RQ4-3, I further identified evaluators' attitudes toward the ML (section 8.5.4.3), how they dealt with (dis)agreements with the ML (section 8.5.4.4), and what they perceived as the ML's limitations (section 8.5.4.5).

8.5.4.1 How did evaluators use the problem timeline (i.e., ML-inferred problem encounters)?

From the interview results, I found that evaluators used the problem timeline in four main ways. First, they used the problem timeline of ML-inferred usability problem encounters as **overviews** to get a sense of the number of potential problems and their distribution over the session even before playing the session. This overview information can be useful for evaluators to get mentally prepared: *“Before the video starts, I looked at the chart to give me a heads up.”*-P39. In the case of the third video where ML identified 17 problems, evaluators used this information to look out for *“big, overarching issues, instead of small little things.”*-P24.

Second, evaluators used the problem timeline as **reminders** or **for anticipations**. It was common that they might zone out while watching or listening to a long-recorded test session, especially when hearing a long period of verbalizations of procedures that do not reveal any problem. In contrast, with the problem timeline, the “spikes” acted as reminders to pull them back and alert them to get ready. *“I'm using the spikes as anticipation...of when I should pay more attention.”*-

P12. *“I’ll be like...a problem’s coming up, and then I’d pay attention, and I will be waiting for the problem to pop up.”-P26.*

Third, evaluators used the “spikes” on the problem timeline as **anchors** to facilitate their re-visitation. *“Then in the second [pass], I wanted to see all the ones that the machine learning highlighted [to] find things that I didn’t notice on my first pass...I just would click where it starts going up, and then go through each one.”-P21.* They also used it for grabbing representative quotes from users *“If I need to grab a quote, I will fast-forward to that part [the ‘spikes’].”-P12.*

Fourth, evaluators used the ‘spikes’ on the problem timeline acted as **guides** to help them **better allocate their attention** during their analysis. Some reported that they paid attention to all areas of the session but paid extra attention to the ‘spikes.’ In contrast, because the ‘spikes’ were visually salient, some paid more attention to the non-spike areas in their first pass of reviewing the session to catch any problems that ML might have missed. *“I should pay attention...when there’s a long flat line...maybe they didn’t pick up something. So I was listening to that part as well.”-P20.*

8.5.4.2 How did evaluators use the feature timeline (i.e., ML’s input features)?

Evaluators in the VisTA condition had access to the feature timeline that visualizes the main input features. But they usually allocated less attention to the feature timeline than the problem timeline. Evaluators mentioned that there was a learning curve to digest and leverage the features and thus typically only considered the feature timeline in the second or third video session when they became relatively familiar with the interface.

Evaluators felt that knowing the input features were helpful because this information allowed them to know *what features were omitted by ML*. Also, it allowed for them to better understand where ML could have missed problems, if the cues for a problem were primarily from the features that ML did not consider, such as visual cues. As a result, they could pay more attention to these features, which in turn allowed for better leverage of ML in their analysis.

In addition to employing the feature timeline to help better understand ML, some evaluators *used the features directly* in their own analysis. Among the features, categories were used more frequently as some observed that “*Observation...could be a potential problem,*”-P13 but “*Reading [is] probably not so much of an issue.*”-P17. On the contrary, evaluators had different opinions about the pitch. Some thought it was helpful; for example, the high pitch could reflect that the user was confused and raising a question. But others thought it was not a reliable signal without understanding the user's normal speaking behaviour. For example, some people tend to raise their tones toward the end of a sentence even if it is not a question.

In contrast, evaluators in the VisTASimple condition, who did not have access to the feature timeline, were asked if they had developed some understanding of the features that ML might have picked up. While many did not have any idea, some pointed out that ML might have used keywords or visual cues (e.g., how much movement the user had). These guesses were either only partially correct or incorrect at all, which could prevent them from using the strategies that evaluators in the VisTA condition used.

8.5.4.3 What were evaluators’ attitudes toward ML-inferred usability problem encounters?

Evaluators developed different perspectives on ML-Inferred usability problem encounters from their user experience. Four evaluators considered ML as a *colleague* or *coworker*, who could provide a *second perspective* on the identified problems. “*It might be picking up something that I had not been thinking about in a different sense...Could it be revealing something else I'm not picking up? Because I have my own confirmation bias.*”-P33.

Two evaluators treated ML-inferred problem encounters as a *backup* when the inferences agreed with them, which increased their confidence in the problems that they identified. “*ML will back up my judgment, helped me confirm that there is a problem.*”-P17. Three evaluators saw ML-inferred problem encounters as *aids* that helped them identify problems faster, not necessarily providing a different perspective that prompted them to reassess their disagreements. “*Use it for*

anticipation. When there is a prediction, I picked out the problems faster. I don't consider it a different perspective.”-P13.

Additionally, four evaluators considered that there was a competition between the evaluators and the ML. Evaluators had this feeling that they wanted to prove that they can do a better job and they had skills that ML may not necessarily possess. *“I didn't feel like it was smarter than me.”-P36, “I want to feel I have skills too.”-P17.*

On the other hand, three evaluators expressed concerns that using ML-inferred problem encounters might cause them to be overly reliant on it and get lazy in their analysis. *“If you don't care about your job you will just follow the chart...Someone still has to watch it (the video).”-P24.*

8.5.4.4 How did evaluators deal with agreement and disagreement with ML-inferred usability problem encounters?

Evaluators felt that the agreement with ML-inferred problem encounters acted as confirmation and reassured the evaluators that they were correct with the identified problems. *“If I find a problem and the model also finds it, I feel more confident”-P26.* Evaluators also felt that seeing the agreement would make them *“pick up the problems faster”-P13.*

Evaluators generally understood that it was possible that ML-inferred problem encounters were imperfect (*“Computer is not perfect...I don't expect it to be”-P17*) and that ML can pick up different problems than they would. When it came to the disagreement, they considered false positives and false negatives of ML-inferred problem encounters differently. When the ML suggested a problem, they generally gave it a second thought even if it might be a false positive. *“I often wonder if I missed any problems, so it is safe to assume there is one (if the ML detects one)”-P21; “It is not a big deal when ML says there is a problem; I examine it and see nothing there.”-P39.*

In contrast, if they thought that the think-aloud user encountered a problem, but the ML did not point it out, they generally considered that the ML missed the problem and would more likely choose to trust themselves. *“By the third session, I started to really believe that the machine was*

just purely picking up more of the audio than the visual. So I think that's why...I gained a little bit more confidence.”-P26. In addition, they generally valued recall over precision. This can be explained by the fact that the goal for evaluators is to find potential problems. Therefore, it is safer to be overly inclusive than omitting potential usability problems.

It is also worth noting that evaluators in VisTASimple generally put less weight on the ML's predictions than those in VisTA when disagreements happened, which is probably because the ML in the VisTASimple, where the input features were not shown, was perceived more like a “black box.” *“I don't know much about what it is based on and how developed the machine learning is, so I don't know how much I can trust it.”-P18.*

8.5.4.5 What did evaluators perceive as the limitations of ML-inferred usability problem encounters?

Evaluators pointed out a number of limitations based on their usage of ML-inferred problem encounters. First, they noted that the ML was often able to detect the moments when the user exhibited symptoms of a problem but did not pinpoint the start and end of the problem. However, observing the problem build-up process was important to fully understand it. *“There was...what I call...a lagging factor. I would have liked to see some of those issues highlighted earlier than some of these spikes on the timeline.”-P14.*

Second, evaluators mentioned that ML did not understand the nuances in a user's personality. For example, some users may prefer to say negative words even when they did not experience too much of a problem. *“I don't think the computer will pick up nuanced behaviours and personalities.”-P17.* Another example is the use of sarcasm, which is hard for the ML to detect based on the text.

Third, they felt that ML-inferred problem encounters did not reflect the context of what users were doing. For example, the ML had difficulty to get the repetition in actions: when users did something repetitively, it could be a problem, but the ML did not seem to pick it up. Additionally, they felt that ML did not consider the number of steps that users took to complete a task as a factor

when accessing problems. For example, taking more steps than needed could mean a problem even if the user completes the task successfully. Lastly, they believe that ML did not fully comprehend the structures of the tasks (e.g., what (sub)tasks did users struggle with?).

8.6 Discussion

Based on the findings of the study, I further discuss how evaluators used and perceived ML-inferred usability problem encounters in their analyses.

8.6.1 The Effect of *ML-inferred Usability Problem Encounters* on the Evaluators' Analysis

Evaluators identified significantly more problems when using VisTA than the Baseline for the third session, but not the first two sessions (see Section 8.5.3.1). One possible reason could be that as this was the first-time evaluators had access to ML, they needed time to learn and understand how to leverage the ML-inferred problems in their analysis over the sessions. Evaluators mentioned that they either did not have much time to carefully consider the problem timeline or were still testing it in the first session. But over time, they were able to develop four general strategies (see Section 8.5.4.1) to use the problem timeline. These strategies encouraged evaluators to be more cautious about their analysis, which was evident by the fact that evaluators using VisTA or VisTASimple paused the videos significantly more than those using Baseline (Section 8.5.2.2 Session Review Strategies). Another possible reason for non-significance in the first two sessions could be that these two sessions were shorter than the last session (see Section 8.2.1) and also contained fewer problems than the last session, and thus the variations between conditions would also be smaller.

Intuitively, an evaluator pointed out, “*Without ML, it is much easier to ignore and let go some issues.*”-P21. When evaluators were watching a session to understand the development of a problem, a new problem might come up, which could take their attention away. If they did not rewind or pause the video in time, they could have missed the locations where they would otherwise want to follow up later. In contrast, the problem timeline acted as an overview, guides,

anchors or anticipations, which facilitated evaluators with pinpointing the areas that they wanted to rewind and pause. This was more effective than using the Baseline to check the points that they might have missed. It is worth pointing out that the way in which evaluators used the problem and feature timelines is inherently tied to their session reviewing behaviour (e.g., pausing and rewinding), and is eventually tied to the number of problems that they identified. The significant difference in the number of problems identified and in the amounts of pausing and rewinding suggests that an ML-enhanced visual analytics tool is capable of helping evaluators become more cautious of their analysis and notice problems that they might have missed.

The evaluations also show that the evaluators in the VisTA condition did not rewind the videos more often than those in the Baseline condition. In contrast, the evaluators in the VisTASimple condition rewound the videos more often than those in the Baseline condition. One potential explanation is that the added problems were visualized at the corresponding timestamps on top of the problem timeline for the evaluators in the VisTA condition (Figure 49). The visualization of these identified problems could also act as “anchors” in addition to the ML-inferred problem encounters on the problem timeline, which might have helped evaluators better determine where they would want to rewind the video if they decided to revisit the video. In contrast, the evaluators in the VisTASimple condition did not have access to the visualization of their identified problems. Consequently, they had no clue where in the think-aloud sessions they had already identified problems and might have a higher chance to rewind the video to points where they had already added problems. This could have increased their need for rewinding more times to locate an area where they might have neglected in their initial analysis.



Figure 49. The *problem timeline* with evaluator-identified problems visualized on top of it. Both the ML-inferred problems and the evaluator identified problems can act as "anchors" to facilitate the re-visitation.

Although evaluators found more problems when using VisTASimple than Baseline for all three sessions, One-way ANOVA did not find a significant difference. This could potentially suggest that having access to the feature timeline that is only available in VisTA in addition to the problem timeline might play a role in encouraging evaluators to identify more problems. One potential reason could be that because the evaluators in the VisTA condition knew what features were considered by the ML, they could better infer when the ML would make a mistake and focus on the features, such as the visual cues, that the ML did not consider. Another potential reason could be that evaluators leverage the feature timeline as additional information in their own analysis instead of merely using it to understand ML. However, individual differences between the Baseline and the VisTASimple could also come into play, as the number of evaluators in each condition is relatively small.

8.6.2 Attitudes toward ML-Inferred Usability Problem Encounters

“Evaluator effect” refers to the fact that different evaluators might identify different sets of problems when analyzing the same session [43]. Although it is recommended to have more than

one UX evaluator analyze a usability test session to reduce potential evaluator effect, fewer than 30% UX practitioners actually had an opportunity to work with others to analyze the same usability test session [34]. The study reveals that a common attitude toward ML-inferred problem encounters was to treat the ML as a “colleague” or a “coworker”, who can provide a second perspective on their analysis or back up their identified problems. This finding points out an opportunity to leverage the ML-inferred problem encounters to help reduce the evaluator effect for UX practitioners, who often operate under resource and time constraints. Toward this goal, I have identified three factors to consider when designing a user interface that leverages the ML-inferred problem encounters to offer a second perspective to UX practitioners.

First, evaluators felt that knowing the severity of the predicted problems can help them to prioritize their analysis especially when they are under time pressure to analyze a large number of test sessions.

Second, evaluators also felt that knowing the confidence level of ML in its inferred problems can also be helpful. For example, they could filter out the low-confident ML-inferred problems and focus more on the high-confident ones, especially when the session is long, and there are many ML-inferred problems.

These two factors raise interesting technical challenges regarding how to automatically detect the severity of problems and enhance the confidence of ML predictions. It also raises an interesting design challenge about how to visualize this information in the same view that is informative but not overwhelming.

Third, evaluators felt that ML would be more like a “colleague” if it could provide explanations for the identified problems. But what kind of explanations are appropriate? And how should they be generated? Although recent research has explored methods for automatically generating explanations [28], some also suggest that the taxonomy for explaining ML to designers is likely *“to be radically different from ones used by data scientists”* [102]. In fact, the feedback from the evaluators who used VisTA echoed the suggestion. Evaluators felt that the current terms used for

input features were too system-orient, making it hard to interpret their meanings. They would prefer these features to be expressed using layman terms, such as the level of surprise, excitement, or frustration.

8.6.3 Reliance on ML-Inferred Usability Problem Encounters

Three evaluators expressed the concern that this may make UX practitioners rely on ML-inferred usability problem encounters too much, thus less diligent in their jobs. However, the study did not find any evidence to support this. First, in all three conditions, evaluators identified problems that ML did not identify, and there was no significant difference between conditions. Similarly, in all the conditions, evaluators disagreed on some of the ML-inferred problem encounters, and there was no significant difference between conditions. This result suggests that evaluators did not just focus on the ML-inferred problem encounters, and also did not just take the words from the ML without scrutinizing them in the VisTA and VisTASimple conditions. Additionally, some evaluators even felt that there was a competition between them and the ML, making them subconsciously eager to prove that they could identify more problems. It is, however, worth noting that as our study duration was short, no baseline trust with the ML had been established. Consequently, it is hard to determine whether evaluators would become over-reliance on ML or develop sustainable cooperative strategies in the long run.

Although none of the evaluators solely relied on the ML-inferred problem encounters without putting in their own thought during analysis, I identified two ways in which evaluators wanted the ML-inferred problem encounters to be presented. One way is to allow evaluators to analyze a test session by themselves in the first pass and then revealing the problem timeline to them in the second pass. In this way, the problem timeline would mainly help them confirm their judgment or double check if they might have missed any problem. The other way is to show the problem timeline all the time. The rationale for this design is that the two-pass reviewing process might not be practical especially when the session is long. This was evident that there were fewer evaluators who adopted the two-pass strategy in the third video, which was the longest among all (Table 24). Although offering evaluators an option to turn on and off the problem timeline seems to be a

compromised approach, it remains an open question of how and when to best present ML to evaluators.

8.6.4 Trust in ML-inferred Usability Problem Encounters

I did not explicitly measure evaluators' trust in ML-inferred usability problem encounters. However, the qualitative analysis on the interview data identify two factors that could have affected their trust in ML-inferred usability problem encounters, including the *sophistication* and the *amount of disagreement*.

The sophistication of ML is determined by the number of features that it considers (e.g., audio and visual features) and whether it understands the context of the task (e.g., the number of steps required to complete a task; meaningless repetitive user actions) or the personality of the user (e.g., the speaking behavior). Evaluators in all three conditions were fairly confident in the problems that they identified no matter how many problems they missed. This could suggest that UX practitioners might suffer from “confirmation bias” [69]. Confirmation bias can be mitigated by revealing the prior probability or input attributions [13,60]. For example, it might be helpful to show the prior probability of catching all the problems from a test session for an average evaluator (e.g., 70%). In this way, evaluators would probably be more willing to consider ML’s inferences when it comes to a disagreement with ML.

The goal of having ML's support is to encourage evaluators to scrutinize their analysis with the input of a different perspective from ML. It is, however, not to overly convince evaluators to agree with ML as it is still an open question whether increasingly agreeing with ML is beneficial for UX analysis. Another way could be to redesign the user interface to prompt evaluators to enter the features that they have considered and then ML could point out the features that they might have neglected. However, how to best design such systems that both deliver ML results and facilitate trust remains to be explored.

8.7 Summary

In this chapter, I took the first step to explore how UX practitioners would use, perceive and react to machine intelligence (i.e., ML-inferred problem encounters and ML's input features) when analyzing recorded think-aloud sessions. I iteratively designed a visual analytics tool, VisTA, that presents the ML-inferred usability problem encounters as a series of “spikes” on a timeline (i.e., *problem timeline*) and the ML's input features within a short time window around the current time in the recorded session (i.e., *feature timeline*) among other annotation and filter functions to facilitate UX practitioners with their analysis. I designed and conducted a three-session between-subjects controlled laboratory study to systematically understand how UX practitioners would leverage ML-inferred problem encounters, ML's input features, and other features to identify usability problems. In addition to demonstrating that UX practitioners identified significantly more problems when using VisTA than Baseline by the last session, the results have also provided deep insights, both quantitatively and qualitatively, about how practitioners leveraged, perceived, and reacted to the ML-inferred problem encounters and the ML's input features.

Based on the analysis of the quantitative and qualitative data from the study, I identified four attitudes that evaluators had toward ML-inferred problem encounters when reviewing think-aloud sessions. They treated the ML as a “colleague,” who can provide a second perspective on their analysis, as a “backup,” that can boost their confidence in their identified problems, as an “aid,” that can simply help them identify problems faster, or as a “competitor,” who motivates them to prove that they can do a better job than the ML.

I identified three ways that evaluators leveraged the problem timeline of the ML-inferred problem encounters. They used the problem timeline as “overviews” to gain a quick understanding of the problem distribution (e.g., which areas have relatively more problems and which areas have relatively fewer problems) even before playing the session recording, as “reminders” to signal themselves when they should be alert and avoid zooming out, as “anchors” to help them better determine where they should revisit, or as “guides” to better allocate their attention between different areas of a recorded think-aloud session.

In addition, I also identified two ways that evaluators used the feature timeline of the ML's input features. They used the feature timeline of ML's input features to better *understand what features that the ML considers and when the ML might make mistakes*, and to use the ML's input features directly as *an extra bit of information* beyond the session recording and ML-inferred problems to help their analysis.

Furthermore, the findings also reveal the evaluators' attitudes toward the agreement and disagreement with the ML-inferred problem encounters. Specifically, they treated the agreement as *confirmations* that reassured them that they were correct with their identified problems. In terms of disagreement, they treated false positives (*i.e.*, the ones that the ML flagged as problems, but they did not think so) and false negative (*i.e.*, the ones that they thought as problems, but the ML did not flag) differently. They felt that false negative is worse than false positive because missing a true problem (*i.e.*, false negative) is more costly than spending effort checking a falsely flagged problem (*i.e.*, false positive). In other words, the evaluators valued the recall over the precision of the ML-inferred usability problem encounters.

Few evaluators worried that UX practitioners might become overly reliant on the ML-inferred usability problem encounters; however, the findings do not support this. In all the three conditions (*i.e.*, VisTA, VisTASimple, Baseline), evaluators identified problems that the ML did not detect and the difference between conditions was not significant; and evaluators in all conditions disagreed on some of the ML-inferred problem encounters and the difference between conditions was not significant either. These results suggest that evaluators put in their own thought into the analysis in all conditions and scrutinized the ML-inferred usability problem encounters when they were available.

Additionally, the findings also show that the sophistication of the ML and the amount of disagreement between the ML and the evaluators seem to affect the evaluators' trust in the ML-inferred usability problem encounters. The evaluators felt that the sophistication of the ML are affected by the number of features that it considers, whether it understands the context of the tasks (e.g., the steps to complete the tasks and therefore any redundant steps performed) and the

personality of the think-aloud user (e.g., the user's typically speaking patterns). This suggests a number of ways that ML can be improved. For example, future work can examine how to leverage multiple modalities of information (e.g., audio, visual, physiological) to make better inferences; how to understand the task composition and the progress of the user in the task; and how to understand the user's typically behavior patterns and use them to adjust the inferences for different users. In terms of the amount of disagreement, the evaluators in the study were all fairly confident in the problems that they identified and therefore tended to trust the ML less if there were more disagreements between their identified problems and the ML-inferred problem encounters. However, this could suggest that the evaluators might have suffered from the "confirmation bias." Future work should examine ways to communicate the potential confirmation bias to evaluators and make them better leverage the ML-inferred problem encounters to help them catch the problems that they might have overlooked.

In sum, through the design and evaluation of VisTA, I demonstrate the promise that UX practitioners can work with and benefit from ML-inferred usability problem encounters, and the quantitative and qualitative results shed light on how UX practitioners perceived, leveraged and integrated the ML-inferred problem encounters and ML's input features into their own analysis flow. The results also highlight potential future research directions to further improve the UX practitioner-ML symbiosis working relationship.

Chapter 9 Conclusion and Future Work

In this chapter, I summarize the key takeaways, reiterate my contributions, discuss the limitations and present future research directions.

9.1 Summary of the Key Takeaways

I have presented my thesis statement in the Introduction chapter (Chapter 1), which is as follows:

Subtle verbalization and speech patterns tend to occur when users encounter problems in concurrent think-aloud sessions; these subtle patterns can be used to automatically detect usability problem encounters, which can be used by UX practitioners as overviews, aids, reminders, anchors, and guides to identify usability problems more effectively.

Of the four contributions that I have made in this dissertation, one extends the understanding of current practices and challenges of conducting and analyzing think-aloud sessions among UX practitioners around the world and grounds this dissertation research. The other three contributions together support this thesis statement. I reiterate these three contributions and present the key takeaways of this dissertation.

9.1.1 Subtle Verbalization and Speech Patterns Tend to Occur When Users Encounter Usability Problems in Concurrent Think-Aloud Sessions

I systematically studied the relationship between users' verbalizations and speech features and usability problems in concurrent think-aloud sessions via three studies (Chapters 4, 5, and 6), each addressing the limitations of its previous one. The findings of the three studies demonstrate that certain patterns of verbalization and speech features act as telltale signs of usability problems in concurrent think-aloud sessions. Segments labeled as the *Observation* category were most likely associated with usability problems. Segments labelled as the *Procedure* category that also contain a description of repeated actions were likely associated with usability problems. Segments labelled as the *Reading* category that last for a long period of time were also likely associated with usability problems. On the contrary, segments labelled as the *Explanation* category were relatively rare and

did not have a clear relationship with usability problems. The findings further show that evaluators often identified problems using combinations of verbalization categories since category combinations were helpful in providing contextual information as to why users were encountering problems. Furthermore, pairs of verbalization categories that contained the *Observation* category were generally more likely associated with problems than those without the *Observation* category.

The findings from Study 2 (Chapter 5) show that the F-measure of using the *Observation* category to locate usability problems was around 0.5. To increase the chance of locating a problem, *sentiment* and *speech features* should be considered in conjunction with the category information. For example, when experiencing problems, users tended to use *negations*, *verbal fillers*, words indicating *uncertainty*, *repetitions*, or *questions*. Therefore, the sentiment of these verbalizations was often *negative*. Furthermore, users tended to verbalize their thought units in *high or low pitches* or with *low speech rates* but rarely changed the loudness of their voices when experiencing problems.

The results of Study 3 (Chapter 6) further demonstrate that the findings of Study 2 are largely generalizable to three factors: the *types of test products* (i.e., physical devices vs. digital systems), and the *modality of the recorded think-aloud sessions* (i.e., audio vs. video recording) and the *visualization of the verbalization and speech features* that evaluators were provided with. The implication is that the same set of verbalization and speech patterns can be used to identify problems that users were experiencing when thinking aloud regardless of whether a physical device or a digital system was used. Usability evaluators can rely on verbalizations alone to identify problems by and large, although certain cues in video streams have additive values to their analysis, such as facial expressions and body language. However, whether these visual cues are consistent across users for locating problems remains to be examined. Moreover, the video stream of a think-aloud session can be informative when the think-aloud user remains silent or frequently uses demonstratives (e.g., this, that) or adverbs of place (e.g., here, there), which makes it difficult to infer what the user is referring to from the audio stream alone. As a result, in such situations, it would be preferable to draw evaluators' attention to the video stream.

9.1.2 Subtle Verbalization and Speech Patterns Can Be Used to Build Effective ML Models to Detect Usability Problem Encounters

Fast-paced analysis for recorded think-aloud sessions is needed to alleviate the burden of usability evaluators. In this dissertation, I took the first step to design and evaluate methods to automate the detection of usability problem encounters in think-aloud test sessions (Chapter 7). The evaluations show that when using *the verbalization and speech features* (i.e., category, sentiment, question, negation, abnormal pitch, and abnormal speech rate) that are shown to be linked to the usability problems as the input, the performance of four different ML models improved compared to only using the basic transcript features (i.e., TF-IDF or word embedding) as the input.

Furthermore, the evaluations show that it is possible to build an ML model for a *product* using its existing users' data to detect the usability problems encountered by a new user; it is also possible to build a user-dependent ML model for a *user* to detect usability problems encountered by the user when she uses a new product. The evaluations also suggest that the *types of tasks* (e.g., guided tasks with prescribed steps to complete and guideless tasks without prescribed steps to complete) that users perform during think-aloud sessions may also affect detection performance. Future work should examine whether *the type of tasks used in think-aloud sessions* and *the difference in user's verbalization behavior* affect the detection of usability problem encounters. As the first step toward automating the usability problem detection, this dissertation work focused on leveraging users' verbalization and speech features to detect usability problem encounters and set a baseline for future exploration.

9.1.3 ML-Inferred Usability Problem Encounters Can Assist UX Practitioners with Identifying Usability Problems More Effectively

I took the first step to explore how UX practitioners would use, perceive and react to ML-inferred usability problem encounters when analyzing recorded think-aloud sessions (Chapter 8). To do so, I designed a visual analytics tool, *VisTA*, that presents *ML-inferred problem encounters as a series of "spikes" on a timeline* and the ML's input features within a short time window around the current time among other functions to facilitate UX practitioners with their analysis. I conducted

three-session between-subjects laboratory study to compare *VisTA*, with *VisTASimple* and *Baseline*. Results show that UX practitioners identified significantly more problems when having access to ML-inferred problem encounters and ML's input features than without having access to such information.

Based on the analysis of the quantitative and qualitative data from the study, I characterized four strategies that UX practitioners used for reviewing think-aloud sessions. Moreover, I provided insights into how they leveraged ML-inferred problem encounters (e.g., as overviews, reminders, anchors, aid and guides) and ML's input features (e.g., as a means to understanding what ML considers and omits or as additional pieces of information) in different conditions.

I also conducted an in-depth investigation on how UX practitioners work with ML in various aspects, such as dealing with agreements and disagreements, limitations of ML, and their reliance and trust for ML. These findings demonstrated the promise that UX practitioners can work with and benefit from ML-inferred problem encounters and ML's input features; the findings also highlighted potential future research directions to further improve the UX practitioner-ML symbiosis working relationship.

9.2 Limitations and Future Research Directions

Sandy Pentland and many other researchers have shown the promise of *honest signals* in predicting many *behavioral and social phenomena* [83], which inspired my dissertation work. Similarly, I hope my dissertation work and the potential future directions can inspire others to further uncover a richer set of *subtle patterns* in verbalization, speech, gaze, facial expression, body language, and physiological measures (e.g., heartbeat, skin conductance response) that are indicative of users' *negative experiences* (e.g., problems, confusions, frustration) as well *positive ones* (e.g., excitement, happiness, satisfaction).

Toward this goal, I highlight the limitations of this dissertation research and discuss potential future research directions.

9.2.1 Further Validating the Findings with *More Products, More Participants, and Evaluators with Different Levels of Familiarity with Test Products*

In this dissertation, I have used different sets of test products, different pools of think-aloud participants, and different sets of usability evaluators for the three studies (Chapters 4, 5, and 6) to better evaluate the validity and generalizability of the subtle verbalization and speech patterns that tend to occur when users encounter problems. The number of test products, however, is still relatively small compared to the ever-growing number of physical and digital products that are available in the world. It would be valuable to *replicate the research with more physical and digital products to further examine the findings*. Although I recruited UX professionals who were working in the industry as evaluators in the three studies, many of the evaluators were graduate students majoring in UX. Thus, the overall experience of the evaluators in the three studies (i.e., Chapters 4, 5, and 6) is relatively less compared to usability evaluators who have worked in the industry for years. It would be valuable to *examine whether and how the years of experience in conducting think-aloud tests might affect the findings of this research*.

The facilitators in the three studies (i.e., Chapters 4, 5, and 6) informed the usability evaluators about the products that the think-aloud users used before they started to evaluate the recorded sessions. Specifically, the facilitators showed the test products, described their main functions, and the tasks that the think-aloud users worked on. Although this introduction provided information about the products that evaluators would evaluate in the recorded think-aloud sessions, I did not provide a chance for the evaluators to use the test products themselves prior to evaluating the recorded think-aloud sessions. I designed the studies in such a way so that the evaluators could identify usability problems that the think-aloud participants experienced in the recordings without being primed by their own experience of using the products. In practice, usability evaluators may have access to the test products and previous research suggests that double experts with knowledge in both usability evaluation and the specific domain might yield better insights [70]. Thus, it would be interesting to explore further *whether having usability evaluators use the test products prior to evaluating think-aloud sessions would have any effect on the finding of this research*.

9.2.2 Understanding the Impact of *Language* and *Age* on the Subtle Verbalization and Speech Patterns that Indicate Usability Problems

In this dissertation, I focused on discovering verbalization and speech patterns in the *English* language. Specifically, I recruited native English speakers who lived in a large metropolitan area in North America to participate in the think-aloud sessions of the three studies that were described in Chapters 4, 5, and 6. In practice, native English speakers may have their own accents, such as American, Australian, British, and Canadian accents. Even within one type of English accent, there are many regional dialects. Each dialect has its own variations in terms of pronunciation and intonation. Furthermore, there are people who speak English as their second language. Therefore, future research should examine *to what extent the findings might be affected by the accents, dialects, and fluency of English if the patterns were used to detect usability problems with users from those geographical regions.*

Furthermore, different languages have different pronunciations and grammars to organize and communicate thoughts and are influenced by different cultures. For example, a field study of think-aloud testing in seven companies in three different countries (i.e., Denmark, China, and India) suggested that the way usability problems are experienced by test participants can be different [23]. Similarly, Shi conducted a field study with companies located in the industrial areas in China and found that Chinese participants tend to have difficulty verbalizing their higher levels of thinking, which might be due to the Chinese holistic thinking style [89]. Thus, if these subtle patterns were to be used for other languages, it is necessary to *examine whether the subtle verbalization and speech patterns that tend to occur when users encounter problems are affected by the languages and the cultures in which the users live.*

Lastly, the think-aloud sessions in the three studies of this dissertation research were conducted with *young adults*. Specifically, the age range of the participants in the three think-aloud studies (Chapters 4, 5, and 6) were 19-26. Language is dynamic, shaped by culture, and has been constantly evolving. Consequently, not all generations speak a language the same as their parents or grandparents. For example, recent research has suggested that age might have an influence on

think-aloud usability testing in terms of task performance and efficiency [78,93]. As a result, one interesting research question is to study *whether and how the subtle verbalization and speech patterns that tend to occur when users encounter problems are affected by different age groups, such as older adults.*

9.2.3 Understanding the Impact of Alternative *Verbalization Categorization Strategies* on the Subtle Patterns that Indicate Usability Problems

I adopted the verbalization categorization strategy that was widely used and cited by previous work [24]. The categorization strategy divides verbalizations into four categories: Reading, Procedure, Observation, and Explanation. Previous research has developed other categorization strategies, which has introduced in the section 2.5. Some of these strategies divide the *Observation* category into more granular sub-categories. For example, Hertzum *et al.* divided the Observation category into four sub-categories: *system observation*, *redesign proposal*, *domain knowledge*, and *user experience* [40]. In contrast, Zhao *et al.* divided it into three subcategories: *expectation*, *positive experience*, and *negative experience* [110]. One interesting research question is to examine whether these more granular categories would allow for better understanding of the connections between users' verbalizations and their experienced usability problems. More granular sub-categories would allow for detecting more nuanced connections between users' verbalization and their experiences. For example, if we were able to establish connections between users' verbalization and speech features and the *redesign proposal* [40], automatic generation of redesign ideas might become possible, which might be able to inspire UX practitioners to generate even better redesign proposals. With Zhao *et al.*'s categorization strategies [110], it is valuable to explore whether there are potential connections between users' verbalizations and speech features and their *positive experiences*. All in all, with these more granular categorization strategies, it is valuable for future work to *examine different verbalization categorization strategies to better understand more subtle and granular connections between users' verbalizations and their experiences.*

9.2.4 Uncovering Subtle Patterns in *Other Types of Think-Aloud Protocols*

I followed Ericsson and Simon's guidelines when conducting think-aloud sessions and did not probe or intervene during the sessions except reminded users to keep talking when they fell into silence for a long time [32]. Our survey study (Chapter 3) and the literature review, however, show that usability practitioners do not always conform to these guidelines (e.g., [76,89]) and may instead employ alternative protocols (e.g., relaxed think-aloud [41], speech-communication [10]). These alternative think-aloud protocols have received mixed results regarding their impact on task performance and the user's ability to make verbalizations [41,95]. Moreover, when using these alternative protocols, practices vary in terms of the instructions, intervention, and prompts that the evaluators use, which are largely because no universal guidelines exist for conducting these alternative think-aloud protocols. Only recently have some researchers started to study the verbalizations in relaxed think-aloud sessions [40]. However, how users' verbalizations relate to usability problems remains largely unknown. Thus, it is interesting to examine *whether and how using intervention and direct instructions that request a particular type of content during the think-aloud sessions affect the subtle patterns that are indicative of usability problems.*

All protocols discussed so far are variations of concurrent think-aloud protocols. Another type of protocol is the retrospective think-aloud protocol. When using the retrospective think-aloud protocol, participants verbalize their thought processes after they complete the task. Although the verbalizations heavily rely on participants' memory and may suffer from post-task rationalization [49], the retrospective think-aloud protocol does have one advantage, which is that verbalizations do not have a direct interference with participants' thought processes during tasks. It is worth exploring *how verbalization and speech patterns in the retrospective think-aloud protocol are indicative of usability problems. For example, would the Observation category still be the category that is most likely associated with problems? Would the Explanation category still be the least popular category? Would users still tend to verbalize in abnormal pitches and speech rates?*

Another direction is to look at the combination of concurrent and retrospective think-aloud protocols, referred to as the hybrid protocol [3]. Recent research shows that when using a hybrid

protocol, the interpretations given after completing the concurrent think-aloud task helped to identify more problems [33] and provide insights into reasons for difficulties that participants encountered during concurrent think-aloud [66]. Thus, I conjecture that there would be more verbalizations labeled as the *Explanation* category in hybrid think-aloud sessions than concurrent think-aloud sessions, which could cause a difference in the categories and category pairs that are more likely associated with problems. A controlled experiment that compares the verbalization and speech patterns that tend to occur when users encounter problems in concurrent think-aloud and the hybrid think-aloud method is needed to ascertain this conjecture.

9.2.5 Discovering the *Subtle Verbalization & Speech Patterns* that are indicative of *the Severity of Usability Problems*

This dissertation focuses on discovering and leveraging subtle verbalization and speech patterns that tend to occur when users encounter usability problems *in general*. Specifically, in the three studies (Chapter 4, 5, and 6), I asked usability evaluators to identify any problems that users were experiencing. Therefore, the findings of this dissertation research reveal the verbalization and speech patterns that are likely to be associated with all usability problems *in general*. In the three studies, I did not, however, request evaluators to rate the severity of the problems primarily because the amount of workload was already considered to be high for the allocated study time based on pilot studies.

As UX practitioners often determine the severity of usability problems based on heuristics [73] and the observation and performance of study participants, I conjecture that there might be subtle signals in users' verbalization, speech or other honest signals that suggest the severity of the problems that users are experiencing. For future research, it would be worth exploring *whether there are correlations between verbalization patterns and a usability problem's severity*.

9.2.6 Creating a *Fully Automatic Pipeline* to Detect Usability Problem Encounters

The objective of my dissertation is to discover, validate, and leverage subtle verbalization and speech patterns that are indicative of usability problems. Toward this goal, I have decided to

conduct a few steps manually to ensure the quality of the data and therefore the validity of the connections between the subtle verbalization and speech patterns and usability problems. For example, after experimenting with current speech recognition APIs [90], I decided to manually transcribe the think-aloud sessions to ensure the transcription quality. I also decided to manually label the verbalization categories to eliminate the impact of potential mis-labels. As a result, *to create the pipeline for detecting usability problem encounters automatically, these manual steps must be automated with high accuracy*. Toward this goal, I have also designed and evaluated ML models to automatically detect the verbalization categories in Section 7.4.5. Future research can explore other methods to improve the category detection accuracy.

Alternatively, future work can also examine whether usability problem encounters can be detected *without transcribing the think-aloud audio or categorizing the verbalizations*, for example, *by using only acoustic features or a combination of acoustic features with other modality features*.

9.2.7 Leveraging the *Wisdom of the Crowds* to Detect Usability Problems

In this dissertation, I have demonstrated that subtle verbalization and speech patterns can be detected and leveraged to detect usability problem encounters, which can then be used by UX practitioners to improve their analysis efficiency. Alternatively, another approach to reducing the workload of detecting usability problems is to design a framework that allows the crowds who do not have UX expertise to perform a quick analysis on small segments of a think-aloud session first and then synthesize these initial analysis from the crowd to infer usability problems. However, toward this goal, there are several challenges must be addressed. First, crowd-sourcing platforms allocate micro-tasks, tasks that only take a small amount of time to complete, to crowd workers. Think-aloud sessions can be long. Thus, it is challenging to *strategically divide a think-aloud session into small segments so that each segment is a meaningful unit for the crowd to work on*. For example, how to ensure that a segment contains enough information for the crowd workers to determine whether the user is encountering a problem without needing to understand its surrounding context? Alternatively, if surround context is needed to resolve ambiguity in a particular segment, how would such a system figure out the context and present it to the crowd

workers. Second, as segments can reveal think-aloud participants' personal information, *how to ensure the privacy of the think-aloud participants (e.g., the identity of the participants) while at the same time provide enough context information to identify usability problems* is another challenge.

9.2.8 Uncovering Subtle Patterns in *Other Modalities (e.g., Gaze, Facial Expressions, Body Language, and Physiological Measures)* that Are Indicative of Users' *Negative and Positive* Experiences

Usability evaluators in the three studies (Chapters 4, 5 and 6) pointed out several important cues in the video modality of the recorded think-aloud sessions that could be useful for locating usability problems, such as facial expressions and body gestures. It is common that people use facial expressions and body gestures to express their emotions and mental states consciously or subconsciously. *Could these data modalities exhibit certain patterns when users encounter usability problems?* The effectiveness of these cues, as the evaluators commented, may vary across people. For example, some participants had neutral faces throughout the test sessions. Nonetheless, there might still be some common patterns in these modalities that are robustly linked to usability problems. If these patterns do exist, what are they? Future research should further examine *whether patterns exist in other data modalities (e.g., facial expressions, gaze patterns, body language, and other physiological measures, such as heartbeats and Galvan skin response) when users encounter usability problems.*

On the other hand, despite multiple data modalities can provide a richer set of information for UX evaluators to leverage, it can also be overwhelming as they have to attend to and digest multiple data streams in order not to miss any important information. However, the spectrum of attention is limited. Consequently, *how to help usability evaluators leverage multiple modalities of data while at the same prioritize their attention* is an important research question. Usability evaluators in the three studies (Chapters 4, 5 and 6) suggested that cues in different modality can be complementary to each other. For example, when think-aloud users fall into silence or use certain words, such as demonstratives or adverbs of place, that are hard to understand the contextual information, it is worth drawing UX evaluators' attention to other modalities, such as the visual

modality. Future research should examine *ways to best leverage different modality cues to direct UX evaluators' attention appropriately so that they will less likely miss any important cues or be overwhelmed by having to monitor multiple modalities equally diligently.*

9.2.9 Building ML Models that Leverage *Bigger Dataset* and Can Detect *Product-specific* and *User-specific Problems More Effectively*

As the first step toward automating the detection of usability problem encounters in think-aloud usability tests, my dissertation focuses on leveraging subtle verbalization and speech features to detect usability problems in general. As a result, this dissertation research should be viewed as a *baseline* for detecting usability problems via subtle signals that users exhibit in think-aloud sessions. There are at least four directions that future work can improve upon the current work.

First, I used the dataset that was collected in Study 3 (i.e., the generalization study) to build and evaluate ML models for detecting usability problem encounters. This dataset included 64 think-aloud sessions, in which eight participants used two physical devices and two digital websites, and the recordings lasted 384 minutes in total. Although the dataset included multiple users and multiple products, it was yet still relatively small compared to the huge number of products and users available in the world. This could be a reason why the data-hungry models, such as CNN and RNN, did not outperform shallow-learning methods, such as SVM and RF, as the deep learning models have much more hyper-parameters to optimize than the shallow-learning methods. One future direction is to curate a larger think-aloud dataset, which includes a larger number of participants and a more diverse set of products, to reassess the performance of ML models and understand whether deep neural networks can achieve better performance.

The challenge with curating such a larger dataset is that conducting large amounts of think-aloud sessions in a controlled lab environment is labor-intensive and time-consuming. One potential solution is to conduct remote usability test sessions, which do not require users to physically present in a lab. With remote usability test, it is possible to recruit a more diverse set of participants around the world.

Second, this dissertation research has demonstrated that it is possible to build an ML model for *a specific product* using its existing users' data to detect the encounters of usability problems for a *new user* (see Section 7.3.2). The performance of such models, however, can still be improved. Future work should examine *how to build more effective product-dependent ML models*. One possible solution might be to weigh the importance of users' verbalized words based on their frequency because it is possible that there might be similarity in a user's verbalizations when the user uses different products. These common words that the user verbalizes when using different products should probably be weighted less compared to the words that are less frequently verbalized only for a specific product.

Third, this dissertation research has also suggested that the *types of tasks* that users perform in think-aloud sessions might have influenced the ML models' performance (see Section 7.4.3). For example, the *guided-tasks*, which were used for physical devices and provided instruction steps for users, might have resulted in higher levels of similarity in users' verbalizations than the *guideless-tasks*, which were used for digital websites and provided no instructions about how to complete the tasks. Future work should examine *whether the type of tasks indeed affects the detection of usability problem encounters and the subtle patterns that are indicative of usability problems*.

Fourth, the evaluations have also suggested that although ML models can be built for each *user* to determine the potential problems that the user might encounter when using a *new* product, the performance of these user-dependent models varied across users (see Section 7.3.3). Future work should *examine what causes the performance difference among different user-dependent models and build more effective user-dependent models*.

9.2.10 Forging a Symbiosis Relationship between UX Practitioners and the ML/AI to better *Identify* and *Interpret* Usability Problems

My current dissertation aims to identify and leverage subtle verbalization and speech patterns to detect *the encounters of usability problems*. Knowing where in the think-aloud sessions users encounter problems (i.e., the encounters of usability problems) is informative for UX practitioners

to locate the problems. In Chapter 8, through the design, development, and evaluation of the visual analytics tool—VisTA, I have demonstrated that UX practitioners can benefit from an ML/AI agent that detects and presents usability problem encounters by treating the ML as a partner (i.e., colleague), an aid, or a competitor. The ML that I have developed, however, can only predict when the user encounters a problem in a think-aloud session but cannot describe what the actual problem is. Consequently, although VisTA can help UX practitioners locate more problems, the UX practitioners *still need to interpret what the actual usability problems are*.

Because interpreting what actual usability problems are requires subjective assessment and judgment, which requires experience that is accumulated by UX practitioners through training and practice, it can be challenging for the AI/ML to truly understand the nuances in users' data and interpret the problems as UX practitioners do. As a result, I believe that AI/ML would not replace UX practitioners. Instead, future UX design and evaluation would need *a harmony collaboration between UX practitioners and the ML/AI agent so that both parties can contribute to the creation of desirable user experience effectively and efficiently*. In this relationship, AI/ML can speed up the analysis by locating the encounters of usability problems among other information, such as the severity of the problems and their confidence in prediction, and present them to UX practitioners, who can incorporate the insights and further leverage their years of experience to make more informed interpretation and judgment.

Toward this goal, I identify the following directions to better understand and forge a sustainable symbiosis relationship between UX practitioners and ML/AI agents that can detect the encounters of usability problems.

First, the evaluators in the VisTA study (Chapter 8) felt that it would be more informative to know *the level of confusion or frustration (i.e., the severity of the usability problems) and the confidence of ML for the identified problems*. Such information could allow them to better prioritize their attention when time and resource is constrained. This opens many Human-Computer Interaction (HCI) challenges. For example, *what are the connections/links between users' verbalizations, speech features, actions on the interface, gaze, facial expressions, and physiology measures and*

the usability problems? Are there any specific words, acoustic features, gaze patterns, facial expressions, motion trajectories, or other physiological measures that suggest the level of confusions or frustration (i.e., the severity of usability problems)?

In the meantime, this dissertation also highlights related Machine Learning (ML) challenges. If patterns indeed exist in users' verbalizations, speech, gaze patterns, and facial expressions that are correlated with their level of confusion or frustration, *could these patterns be used to build effective computational models that can detect the level of confusion or frustration (i.e., the severity of usability problems)?*

Second, in addition to making ML/AI agents more intelligent and powerful, another key component to forge a successful symbiosis working relationship between UX practitioners and ML/AI agents is to understand how UX practitioners perceive, interact and incorporate ML/AI agents' intelligence. Therefore, the research question is *how to present the rich spectrum of ML-inferred information (e.g., automatically detected problems and their severity levels, and the ML's confidence in its prediction and the input features that it leverages) to UX practitioners to help them effectively and efficiently analyze think-aloud sessions yet without overwhelming them?*

To answer this question, it is necessary to understand the trade-off between presenting rich analytical data and managing the cognitive load of UX practitioners. One potential approach that I learned from the evaluators in the VisTA study is to allow for turning on and off different functions. For example, some evaluators in the study preferred to see the ML-enhanced information in their second pass of the analysis in which they can leverage the ML to check the problems that they might have missed or to confirm the problems that they have identified. However, this solution assumes that UX practitioners know what types of information they want to access at any given time of their analysis process and are willing to do so. Future research should examine whether this assumption holds true and further explore different visual analytical approaches to facilitate UX practitioners to understand and incorporate the ML/AI intelligence.

Third, our study with VisTA also found that some evaluators treated ML/AI as a “colleague” who could provide a second perspective. This is a potential opportunity to help UX practitioners to reduce the evaluator effect [43] by consulting to ML/AI for their “opinion.” However, in order to ML/AI more like a “college,” our study evaluators wanted the ML/AI to be able to explain its decisions. Future work should explore *ways to enable ML/AI to better explain or describe the problems that it identifies*. This would probably require the community to better understand the language that UX practitioners use to communicate problems among themselves as recent research suggested that the taxonomy for explaining ML to designers is likely “to be radically different from ones used by data scientists” [102].

9.2.11 Designing *Real-time Intelligent* Systems to Assist UX Practitioners to *Conduct Usability Test Sessions More Strategically*

In addition to the challenges related to analyzing think-aloud sessions that this dissertation tackles, our international survey study (Chapter 3) also discovered the challenges related to conducting think-aloud sessions. The survey study found that although UX practitioners are aware of the potential dangers of probing users during think-aloud sessions, which includes altering the users’ thought processes, they also feel that they need to probe the participants to understand their thought processes to better understand the problems and their causes, especially when the participants forget to verbalize their thoughts. Understanding when to probe users to minimize the interruptions can be tricky and is an art that requires years of practices to master. *Is it possible to hint UX practitioners, especially those who have relatively less experience with conducting think-aloud sessions, when they should probe users with a question and when they should just wait patiently and let users think aloud without interrupting them?*

This dissertation has identified and validated subtle verbalization and speech patterns that are indicative of usability problems and has further demonstrated that these subtle patterns can be used to build effective ML models to detect the encounters of usability problems. One potential use of the automatically detected usability problem encounters is to *suggest appropriate moments for the UX practitioners to probe participants with further questions* to understand exactly what their

problems are while keeping the overall interactions to a minimal level. To realize this goal, further research is needed to *make possible the real-time processing and prediction based on users' verbalization and speech signals*. Moreover, it is also important to explore *how to deliver the ML-inferred moments for probing to UX practitioners so as not to disturb their observation and participants' thinking aloud process*.

References

1. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, and others. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.
2. 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: 111–132. Retrieved from <http://www.upassoc.org>.
3. Obead Alhadreti and Pam Mayhew. 2018. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–12. <https://doi.org/10.1145/3173574.3173618>
4. 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. <https://doi.org/10.1145/1240624.1240838>
5. Pamela M. Auble, Jeffery J. Franks, Salvatore A. Soraci, Salvatore A. Soraci, and Salvatore A. Soraci. 1979. Effort toward comprehension: Elaboration or “aha”? *Memory & Cognition* 7, 6: 426–434. <https://doi.org/10.3758/BF03198259>
6. Tyra Beer, Tatyana Anodenko, and Andrew Sears. 1997. A Pair of Techniques for Effective Interface Evaluation: Cognitive Walkthroughs and Think-Aloud Evaluations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 41, 1: 380–384. <https://doi.org/10.1177/107118139704100184>
7. André Berthold and Anthony Jameson. 1999. Interpreting Symptoms of Cognitive Load in Speech Input. In *UM99 user modeling*. Springer, Vienna, 235–244. https://doi.org/10.1007/978-3-7091-2490-1_23

8. Martin Blanchard, Nathaniel D’Mello, Sidney Olney, Andrew M. Nystrand. 2015. Automatic Classification of Question & Answer Discourse Segments from Teacher’s Speech in Classrooms. *International Educational Data Mining Society*. Retrieved May 6, 2019 from <https://eric.ed.gov/?id=ED560555>
9. Paul Boersma. 2006. Praat: doing Phonetics by Computer.
10. Ted Boren and Judith Ramey. 2000. Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication* 43, 3: 261.
11. Victoria A. Bowers and Harry L. Snyder. 1990. Concurrent versus retrospective verbal protocol for comparing window usability. *Proceedings of the Human Factors Society* 34, 17: 1270–1274. <https://doi.org/10.1177/154193129003401720>
12. 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 27th international conference on Human factors in computing systems - CHI 09*, 1619–1628. <https://doi.org/10.1145/1518701.1518948>
13. Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, 160–169.
14. Kapil Chalil Madathil and Joel S. Greenstein. 2011. Synchronous remote usability testing: a new approach facilitated by virtual worlds. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*, 2225–2234. <https://doi.org/10.1145/1978942.1979267>
15. Kapil Chalil Madathil and Joel S. Greenstein. 2011. Synchronous remote usability testing: a new approach facilitated by virtual worlds. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11*, 2225–2234. <https://doi.org/10.1145/1978942.1979267>

16. Elizabeth Charters. 2003. The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal* 12, 2: 68–82. <https://doi.org/10.26522/brocked.v12i2.38>
17. Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8, 2: 9:1--9:20. <https://doi.org/10.1145/3185515>
18. Michelene T.H. Chi, Nicholas De Leeuw, Mei Hung Chiu, and Christian Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science* 18, 3: 439–477. [https://doi.org/10.1016/0364-0213\(94\)90016-7](https://doi.org/10.1016/0364-0213(94)90016-7)
19. Parmit K. Chilana, Jacob O. Wobbrock, and Andrew J. Ko. 2010. Understanding usability practices in complex domains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2337–2346. <https://doi.org/10.1145/1753326.1753678>
20. Jason M. Chin and Jonathan W. Schooler. 2008. Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. *European Journal of Cognitive Psychology* 20, 3: 396–413. <https://doi.org/10.1080/09541440701728623>
21. Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Retrieved May 6, 2019 from <http://arxiv.org/abs/1406.1078>
22. Herbert H Clark and Jean E Fox Tree. 2002. Using “uh” and “um” in spontaneous speaking. *Cognition* 84, 1: 73--111. [https://doi.org/http://dx.doi.org/10.1016/S0010-0277\(02\)0](https://doi.org/http://dx.doi.org/10.1016/S0010-0277(02)0)
23. Torkil Clemmensen, Qingxin Shi, Jyoti Kumar, Huiyang Li, Xianghong Sun, and Pradeep Yammiyavar. 2007. Cultural Usability Tests – How Usability Tests Are Not the Same All

- over the World. In *Usability and Internationalization. HCI and Culture*. Springer Berlin Heidelberg, Berlin, Heidelberg, 281–290. https://doi.org/10.1007/978-3-540-73287-7_35
24. Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication* 53, 3: 202–215. <https://doi.org/10.1109/TPC.2010.2052859>
 25. Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*, 278–288.
 26. Joseph S. Dumas and Janice. Redish. 1999. *A practical guide to usability testing*. Intellect Books. Retrieved April 16, 2019 from <https://dl.acm.org/citation.cfm?id=600280>
 27. Steven J. Durning, Anthony R. Artino, Thomas J. Beckman, John Graner, Cees Van Der Vleuten, Eric Holmboe, and Lambert Schuwirth. 2013. Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical Teacher* 35, 9: 720–726. <https://doi.org/10.3109/0142159X.2013.801938>
 28. Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, 263–274. <https://doi.org/10.1145/3301275.3302316>
 29. Fatma Elbabour, Obead Alhadreti, and Pam Mayhew. 2017. Eye tracking in retrospective think-aloud usability testing: Is there added value? *Journal of Usability Studies* 12, 3: 95–110. Retrieved from <https://dl.acm.org/citation.cfm?id=3190864>
 30. Elling Sanne, Lentz Leo, and Menno De Jong. 2012. Combining Concurrent Think-Aloud Protocols and Eye-Tracking Observations : An Analysis of Verbalizations. *Ieee Transactions on Professional Communication* 55, 3: 206–220.

<https://doi.org/10.1109/TPC.2012.2206190>

31. K. Anders Ericsson. 2003. Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness* 10, 9: 1–18.
32. K. Anders Ericsson and Herbert A. Simon. 1984. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.
33. Asbjørn Følstad. 2007. Work-Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work-Support Systems. *International Journal of Human-Computer Interaction* 22, 3: 217–245. <https://doi.org/10.1080/10447310709336963>
34. Asbjørn Følstad, Effie Law, and Kasper Hornbæk. 2012. Analysis in practical usability evaluation: a survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2127–2136. <https://doi.org/10.1145/2207676.2208365>
35. Mark C Fox, K Anders Ericsson, and Ryan Best. 2011. Do Procedures for Verbal Reporting of Thinking Have to be Reactive? *Psychological bulletin* 137, 2: 316.
36. Jill Gerhardt-Powals. 1996. Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction* 8, 2: 189–211. <https://doi.org/10.1080/10447319609526147>
37. Amy M. Gill and Blair Nonnecke. 2012. Think aloud: Effects and Validity. In *Proceedings of the 30th ACM international conference on Design of communication - SIGDOC '12*, 31. <https://doi.org/10.1145/2379057.2379065>
38. Frieda Goldman-Eisler. 1986. *Psycholinguistics: Experiments in spontaneous speech*. Double day, New York, NJ.
39. David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P.N. Rao.

2008. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 835.
<https://doi.org/10.1145/1357054.1357187>
40. 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: 557–570.
<https://doi.org/10.1080/10447318.2015.1065691>
41. Morten Hertzum, Kristin D. Hansen, and Hans H.K. K Andersen. 2009. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology* 28, 2: 165–181.
<https://doi.org/10.1080/01449290701773842>
42. Morten Hertzum and Kristin Due Holmegaard. 2013. Thinking Aloud in the Presence of Interruptions and Time Constraints. *International Journal of Human-Computer Interaction* 29, 5: 351–364. <https://doi.org/10.1080/10447318.2012.711705>
43. Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction* 13, 4: 421–443. https://doi.org/10.1207/S15327590IJHC1304_05
44. Morten Hertzum, Rolf Molich, and Niels Ebbe Jacobsen. 2014. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour and Information Technology* 33, 2: 143–161. <https://doi.org/10.1080/0144929X.2013.783114>
45. T. Hianik, M. Haburcák, K. Lohner, E. Prenner, F. Paltauf, and A. Hermetter. 1998. Compressibility and density of lipid bilayers composed of polyunsaturated phospholipids and cholesterol. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 139, 2: 189–197. <https://doi.org/10.1080/13645579.2011.625764>

46. Masahiro Hori, Yasunori Kihara, and Takashi Kato. 2011. Investigation of indirect oral operation method for think aloud usability testing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 38–46. https://doi.org/10.1007/978-3-642-21753-1_5
47. Kasper Hornbæk. 2010. Dogmas in the assessment of usability evaluation methods. *Behaviour and Information Technology* 29, 1: 97–111. <https://doi.org/10.1080/01449290801939400>
48. Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth international AAAI conference on weblogs and social media*, 216–225.
49. M. D. T de Jong, P. J Schellens, and M. J Van den Haak. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting With Computers* 16, 6: 1153–1170. Retrieved from <https://academic.oup.com/iwc/article-abstract/16/6/1153/769631>
50. Gabriela Jurca, Theodore D. Hellmann, and Frank Maurer. 2014. Integrating Agile and User-Centered Design: A Systematic Mapping and Review of Evaluation and Validation Studies of Agile-UX. In *Proceedings - 2014 Agile Conference, AGILE 2014*, 24–32. <https://doi.org/10.1109/AGILE.2014.17>
51. Patrik N. Juslin and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129, 5: 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>
52. Claire-Marie Karat, Robert Campbell, and Tarra Fiegel. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, 397–404. <https://doi.org/10.1145/142750.142873>

53. Jesper Kjeldskov, Mikael B. Skov, and Jan Stage. 2004. Instant data analysis: conducting usability evaluations in a day. In *Proceedings of the third Nordic conference on Human-computer interaction - NordiCHI '04*, 233–240. <https://doi.org/10.1145/1028014.1028050>
54. Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–14. <https://doi.org/10.1145/3290605.3300641>
55. Emiel Krahmer and Nicole Ummelen. 2004. *Thinking about thinking aloud: A comparison of two verbal protocols for usability testing*. <https://doi.org/10.1109/TPC.2004.828205>
56. Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)* 333: 2267–2273. <https://doi.org/10.1145/2808719.2808746>
57. Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissioti, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior* 32, 4: 195–214. <https://doi.org/10.1007/s10919-008-0055-9>
58. Darryn Lavery, Gilbert Cockton, and Malcolm P. Atkinson. 1997. Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology* 16, 4–5: 246–266. <https://doi.org/10.1080/014492997119824>
59. Clayton Lewis. 1982. *Using the “thinking-aloud” method in cognitive interface design*.
60. Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, 157–166.
61. Kristiyan Lukanov, Horia A. Maior, and Max L. Wilson. 2016. Using fNIRS in Usability

- Testing. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*: 4011–4016. <https://doi.org/10.1145/2858036.2858236>
62. Craig M. MacDonald and Michael E. Atwood. 2013. Changing perspectives on evaluation in HCI: past, present, and future. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, 1969–1978. <https://doi.org/10.1145/2468356.2468714>
 63. Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–12. <https://doi.org/10.1145/3173574.3173922>
 64. 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: 2–19. <https://doi.org/10.1109/TPC.2011.2182569>
 65. Sharon McDonald and Helen Petrie. 2013. The effect of global instructions on think-aloud testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 2941–2944. <https://doi.org/10.1145/2470654.2481407>
 66. Sharon McDonald, Tingting Zhao, and Helen M. Edwards. 2013. Dual Verbal Elicitation: The Complementary Use of Concurrent and Retrospective Reporting Within a Usability Test. *International Journal of Human-Computer Interaction* 29, 10: 647–660. <https://doi.org/10.1080/10447318.2012.758529>
 67. Sharon McDonald, Tingting Zhao, and Helen M Edwards. 2016. Look who’s talking: Evaluating the utility of interventions during an interactive think-aloud. *Interacting with Computers* 28, 3: 387–403. <https://doi.org/10.1093/iwc/iwv014>
 68. Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *11th IEEE International Conference and Workshops on*

- Automatic Face and Gesture Recognition (FG)*, 1–6. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7163127/>
69. Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2: 175–220.
 70. Jakob Nielsen. 1993. *Usability engineering*. Elsevier.
 71. Jakob Nielsen. 1994. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human - Computer Studies* 41, 3: 385–397. <https://doi.org/10.1006/ijhc.1994.1065>
 72. Jakob Nielsen. 1995. 10 usability heuristics for user interface design. *Nielsen Norman Group* 1, 1. Retrieved from [http://courses.ischool.utexas.edu/rbias/2014/Spring/INF385P/files/10 Usability Heuristics for User Interface Design.docx](http://courses.ischool.utexas.edu/rbias/2014/Spring/INF385P/files/10%20Usability%20Heuristics%20for%20User%20Interface%20Design.docx)
 73. Jakob Nielsen. 1995. Severity Ratings for Usability Problems. *54*, 1–2. Retrieved from <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>
 74. Jakob Nielsen and Robert L. Mack. 1994. *Usability inspection methods*. Wiley. Retrieved May 21, 2019 from <https://dl.acm.org/citation.cfm?id=189209>
 75. Jakob Nielson. 2014. Demonstrate Thinking Aloud by Showing Users a Video. *Nielson Norman Group: Evidence-Based User Experience Research, Training, and Consulting*. Retrieved from <https://www.nngroup.com/articles/thinking-aloud-demo-video/>
 76. Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? an explorative study of think-aloud testing. In *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06*, 209. <https://doi.org/10.1145/1142405.1142439>
 77. Kenneth R. Ohnemus and David W Biers. 1993. Retrospective versus concurrent thinking-

- out-loud in usability testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 37, 17: 1127–1131. <https://doi.org/10.1177/107118137902300152>
78. Erica Olmsted-Hawala and Jennifer Romano Bergstrom. 2012. Think-Aloud Protocols: Does Age Make a Difference? In *Stc Technical Communication Summit*.
79. 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 28th international conference on Human factors in computing systems - CHI '10*, 2381. <https://doi.org/10.1145/1753326.1753685>
80. 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 28th international conference on Human factors in computing systems - CHI '10*, 2381. <https://doi.org/10.1145/1753326.1753685>
81. Erica L Olmsted-Hawala, Elizabeth D Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Olmsted-hawala et al., 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 Think-aloud protocols Alterna.pdf. In *Professional Communication Conference (IPCC), 2010 IEEE International*, 60–66.
82. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct: 2825–2830.
83. Alex Pentland. 2009. *Honest Signals: How They Shape Our World*. MIT press. <https://doi.org/10.1145/2072298.2072374>

84. Lloyd R. Peterson. 1969. Concurrent verbal activity. *Psychological Review* 76, 4: 376–386. <https://doi.org/10.1037/h0027443>
85. Matthew F. Pike, Horia A. Maior, Martin Porcheron, Sarah C. Sharples, and Max L. Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, 3807–3816. <https://doi.org/10.1145/2556288.2556974>
86. Jenny Preece, Yvonne Rogers, and Helen Sharp. 2015. *Interaction design : beyond human-computer interaction*. Wiley.
87. Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: how to plan, design and conduct effective tests*. John Wiley & Sons.
88. Klaus R. Scherer. 1995. Expression of emotion in voice and music. *Journal of Voice* 9, 3: 235–248. [https://doi.org/10.1016/S0892-1997\(05\)80231-0](https://doi.org/10.1016/S0892-1997(05)80231-0)
89. Qingxin Shi. 2008. A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In *Proceedings of the 5th Nordic conference on Human-computer interaction building bridges - NordiCHI '08*, 344. <https://doi.org/10.1145/1463160.1463198>
90. Glen Shires and Hans Wennborg. 2012. Web Speech API Specification. *Speech API Community Group, W3C*. <https://doi.org/10.1021/jf001495e>
91. Vicki L. Smith and Herbert H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32, 1: 25–38. <https://doi.org/10.1006/JMLA.1993.1002>
92. Vikrant Soman and Anmol Madan. 2010. Social signaling: Predicting the outcome of job interviews from vocal tone and prosody. In *Proceedings of IEEE International Conference on Acoustic Speech Signal Process*.

93. Andreas Sonderegger, Sven Schmutz, and Juergen Sauer. 2016. The influence of age in usability testing. *Applied Ergonomics* 52: 291–300.
<https://doi.org/10.1016/j.apergo.2015.06.012>
94. Siegfried L Sporer and Barbara Schwandt. 2006. Paraverbal Indicators of Deception: a Meta-Analytic Synthesis. *The Journal of Applied Cognitive Psychology* 20, 4: 421–446.
<https://doi.org/10.1002/acp.1190>
95. S. Tirkkonen-Condit. 2006. Think-Aloud Protocols. In *Encyclopedia of Language & Linguistics*, 678–686. <https://doi.org/10.1016/B0-08-044854-2/00479-X>
96. Zuowei Wang, Xingyu Pan, Kevin F. Miller, and Kai S. Cortina. 2014. Automatic classification of activities in classroom discourse. *Computers & Education* 78: 115–123.
<https://doi.org/10.1016/J.COMPEDU.2014.05.010>
97. Heinz Werner and Bernard Kaplan. 1963. *Symbol formation*. Wiley, Oxford, England.
Retrieved May 4, 2019 from <https://psycnet.apa.org/record/1963-35019-000>
98. Cathleen Wharton. 1994. The cognitive walkthrough method: A practitioner’s guide. *Usability inspection methods*.
99. Richard B Wright and Sharonlyn A Converse. 1992. Method bias and concurrent verbal protocol in software usability testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 36, 16: 1220–1224. <https://doi.org/10.1177/154193129203601608>
100. Jasy Liew Suet Yan, Nancy McCracken, and Kevin Crowston. 2014. Semi-Automatic Content Analysis of Qualitative Data. In *iConference 2014 Proceedings*.
<https://doi.org/10.9776/14399>
101. Jasy Suet Liew Yan, Nancy McCracken, Shichun Zhou, and Kevin Crowston. 2014. Optimizing Features in Active Machine Learning for Complex Qualitative Content Analysis. *Proceedings of the ACL 2014 Workshop on Language Technologies and*

- Computational Social Science* 56, M1: 44–48. <https://doi.org/10.3115/v1/w14-2513>
102. Qian Yang. 2018. Machine Learning as a UX Design Material: How Can We Imagine Beyond Automation, Recommenders, and Reminders? In *2018 AAAI Spring Symposium Series*.
 103. Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*, 585–596. <https://doi.org/10.1145/3196709.3196730>
 104. Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning adaptive mobile experiences when wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 565–576.
 105. Po-Yin Yen and Suzanne Bakken. 2009. A comparison of usability evaluation methods: heuristic evaluation versus end-user think-aloud protocol - an example from a web-based communication tool for nurse scheduling. In *AMIA Symposium Proceedings*, 714–18. <https://doi.org/10.1016/j.ajem.2009.10.015>
 106. Bo Yin and Fang Chen. 2007. Towards Automatic Cognitive Load Measurement from Speech Analysis. In *Human-Computer Interaction. Interaction Design and Usability*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1011–1020. https://doi.org/10.1007/978-3-540-73105-4_111
 107. Bo Yin, Natalie Ruiz, Fang Chen, and M. Asif Khawaja. 2007. Automatic cognitive load detection from speech features. In *Proceedings of the 2007 conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artifacts and environments - OZCHI '07*, 249. <https://doi.org/10.1145/1324892.1324946>
 108. Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus.

The Journal of the Acoustical Society of America 123, 5: 3878–3878.

<https://doi.org/10.1121/1.2935783>

109. 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 - NordiCHI '10*, 581. <https://doi.org/10.1145/1868914.1868979>
110. 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 and Information Technology* 33, 2: 162–182.
<https://doi.org/10.1080/0144929X.2012.708786>
111. Haiyi Zhu, Robert E. Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 3431. <https://doi.org/10.1145/1978942.1979453>
112. oTranscribe. Retrieved September 8, 2017 from <http://otranscribe.com/>