박사학위논문
Ph.D. Dissertation

AI 시스템 설계에서 간과된 사용자 그룹의 사용 패턴과 인식 분석을 통한 추가적인 불이익 탐구

Unveiling the Additional Risks Faced by Overlooked User Groups of AI Systems via Analyzing their Usage Patterns and Perceptions

2025

김서영 (金瑞永 Kim, Seoyoung)

한국과학기술원

Korea Advanced Institute of Science and Technology

박사학위논문

AI 시스템 설계에서 간과된 사용자 그룹의 사용 패턴과 인식 분석을 통한 추가적인 불이익 탐구

2025

김서영

한국과학기술원

전산학부

AI 시스템 설계에서 간과된 사용자 그룹의 사용 패턴과 인식 분석을 통한 추가적인 불이익 탐구

김서영

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2024년 12월 2일

심사위원장	김주호	(인)
심사위원	오혜연	(인)
심사위원	홍화정	(인)
심사위원	Xiaojuan Ma	(인)
심사위원	Joseph Seering	(인)

Unveiling the Additional Risks Faced by Overlooked User Groups of AI Systems via Analyzing their Usage Patterns and Perceptions

Seoyoung Kim

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
December 2, 2024

Approved by

Juho Kim
Associate Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DCS

김서영. AI 시스템 설계에서 간과된 사용자 그룹의 사용 패턴과 인식 분석을 통한 추가적인 불이익 탐구. 전산학부 . 2025년. 99+vi 쪽. 지도교수: 김주호. (영문 논문)

Seoyoung Kim. Unveiling the Additional Risks Faced by Overlooked User Groups of AI Systems via Analyzing their Usage Patterns and Perceptions. School of Computing . 2025. 99+vi pages. Advisor: Juho Kim. (Text in English)

Abstract

Artificial Intelligence (AI) systems are becoming integral to daily life, providing various conveniences to users. However, AI systems are often developed focusing on behaviors or perceptions of the typical users. This thesis investigates the additional risks faced by the user groups overlooked during the design of AI systems. For this, I investigate the overlooked usage patterns and perceptions that are different from that of typical users in various AI systems prevalent in our daily lives. To be specific, the thesis investigates how overlooked differences in (1) usage patterns of older adults while using video recommendation systems, (2) usage patterns of non-native speakers using large language models for writing, (3) perception differences of privacy-wise sensitive users using behavior log-based detection systems, and (4) perception differences of non-native speakers using automatic speech recognition systems may bring additional risks to them. By uncovering the additional risks the overlooked user groups may face, this work calls for a more inclusive AI development that ensures to take into account the diverse usage patterns and perceptions.

Keywords Inclusive AI, AI risks, usage patterns, user perceptions

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Artificial Intelligence (AI) systems are transforming our daily lives, offering greater convenience across various domains such as healthcare, education, entertainment, and communication. For example, Automatic Speech Recognition (ASR) technology improved with the advancement of its models, showing significantly improved transcription accuracy. This improvement has brought ASR to be widely used in various applications in our daily lives, ranging from voice assistants to captions in meeting support technologies, and even to tools for creating captions for accessibility in video content. These applications highlight ASR's growing role in enhancing communication and accessibility in various aspects of our lives.

However, these advancements in AI systems do not bring benefits and convenience to all users. This is because AI has been developing while focusing on the *typical* users. For instance, ASR system has improved a lot in their performance after changing to end-to-end model from the combination of the acoustic model and language model. This change has brought more convenience to the *typical* users (e.g., native speakers using ASR for online meetings) due to the increased performance. However, a non-native speaker who uses ASR systems as a learning tool to identify and correct errors in their speech to improve their English skills may not benefit from this improvement. Due to this change of using end-to-end model, when they use an ASR system and say something that contains a grammatical error, the system automatically generates a transcription that corrects the error, rather than reflecting what was actually spoken. While this behavior makes the generated transcription clearer to be understood for most users, it removes the opportunity for the non-native speaker who is trying to learn from the ASR output about their grammar mistakes. This demonstrates how the advancements in ASR, while beneficial for the typical users, can unintentionally disadvantage certain user groups.

Leaving out certain users with different behaviors and perceptions extends far beyond ASR. As AI systems become more pervasive, the exclusion of certain user groups may widen digital inequalities. The lack of considering diverse users in the process of developing AI systems may undermine the potential for equitable innovation, limiting the societal impact of these systems. Without a clear understanding of who these user groups are and what are the additional risks or challenges do they face, future AI development would be consistently advance in a way that would leave out certain users.

Under the importance of understanding the additional risks or challenges faced by the overlooked user groups, my thesis tries to understand this by analyzing users' usage patterns and perceptions towards the AI systems. Thus, my thesis statement is:

**User groups overlooked during the AI system's design process face additional risks using the AI systems;
understanding their usage patterns or perceptions can reveal these risks.**

The additional risks here refers to the hidden risks or challenges that they may encounter by using the AI systems. This do not necessarily mean the user groups would not be able to use the AI systems. For instance, let's take an example of a non-native speaker using automatic speech recognition systems in a remote meeting. Although the non-native speaker might have used the automatic speech recognition system successfully, if the listeners read the resulting captions and more were able to perceive that the person's accent as 'negatively', then this may be also one of the additional risks that person faced during the usage of automatic speech recognition system.

The user groups facing the additional risks or challenges while using the AI system may differ according to the different characteristics of AI systems. This is because how the user groups use or perceive AI systems differ according to various factors of the AI systems, such as its types, characteristics, or design. As there are diverse AI systems with different characteristics, this thesis investigates the additional risks in diverse AI systems by understanding usage patterns or perceptions of the AI systems that are widely used in our daily lives, such as recommendation systems, large language models, behavioral log-based detection systems, and automatic speech recognition systems.

For each system this thesis identifies different additional risks faced by the overlooked user groups that may not be faced by other user groups by understanding their differences in their usage patterns and perceptions with typical users that have not been considered while designing the AI system. For this, this thesis investigates how overlooked differences in usage patterns may bring additional risks to (1) older adults while using video recommendation systems and (2) non-native speakers using large language models for writing. This thesis also investigates how overlooked differences in perceptions may bring additional risks to (1) privacy-wise sensitive users using behavior log-based personality detection systems and (2) non-native speakers using automatic speech recognition systems.

For this, this thesis utilizes a various methodologies to investigate the additional risks faced by overlooked user groups, such as interviews, surveys, experiments, research-probe, and large-scale log data analysis. Through quantitative and qualitative analysis, this thesis identifies the different usage patterns and perceptions of these users, which eventually unveils the additional risks faced by overlooked user groups of AI systems.

1.1 Thesis overview

This section presents the overview of the thesis.

Chapter 2. Related Work

This chapter reviews existing research relevant to each of the work that will be presented in Chapter 3, 4, 5, and 6. To be specific, this chapter discusses the related work on usage pattern differences in video recommendation systems, usage pattern differences in large language models, perception differences in behavior log-based personality detection systems, and perception differences in automatic speech recognition systems.

Chapter 3: How Older Adults Use Online Videos for Learning

This chapter investigates the motivations, interaction patterns, and challenges of older adults using online videos for learning. Through interviews and log analysis of a Korean MOOC platform, the study reveals how older adults' behaviors in online video platforms differ from those of non-older adults and discusses how their behaviors may give them higher risks for fatigue when recommended with many videos to just increase serendipity.

Chapter 4: Understanding Non-Native and Native Speakers' Use of LLMs in Writing

This chapter examines how non-native speakers (NNS) and native speakers (NS) interact differently with large language models for writing. The study identifies that NNS are more likely to request LLM for a draft while they are more likely to rely on the LLM-genrated draft, which may lead them to decreased writing authenticity for NNS with the current LLM design, which gives freedom of user control while LLM just passively provide responses under the goal to fulfill users' intent.

Chapter 5: Understanding Users' Perception Towards Automated Personality Detection with Group-specific Behavioral Data

This chapter explores how users perceive automated personality assessment (APA) systems that detects user's personality based on their behavior logs. It identifies factors such as privacy concerns, unwanted behavior changes, and trust issues that influence user perceptions. The findings offer insights into designing APA systems that balance accuracy with user's perception, while discussing how those with privacy concerns may face the risks of getting low accuracy results.

Chapter 6: Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker's Accent

This chapter investigates how listeners perceive the same ASR outputs differently depending on whether the speaker is a native or non-native speaker. This chapter examines how the listener's different perceptions can lead to unfair attributions of errors to certain user groups, thus, inferring that an AI system with no performance gap may not mean that the system is fair.

Chapter 7: Discussion

This chapter discusses the generalization of the results and broader implications for future AI systems. It also addresses how to identify the user groups with additional risks and strategies for reducing those risks.

Chapter 8: Conclusion

This chapter summarizes the key findings of the thesis, emphasizing how AI systems can disadvantage overlooked user groups with differing usage patterns or perceptions, and outlines the expected impact of fostering awareness and rethinking design principles to create more inclusive AI systems.

Chapter 2. Related Work

This chapter discusses the related work on (1) usage pattern differences in video recommendation systems and large language models and (2) perception differences in behavior log-based personality detection systems and automatic speech recognition systems.

2.1 Usage pattern difference in video recommendation systems

2.1.1 Learning through Online Videos

Watching online videos is a promising way to pursue learning. It is highly accessible compared to traditional education, with little restriction on time and location of learners [95]. Moreover, most online learning services are more affordable than their offline counterparts, allowing users to easily access them [158].

Although highly accessible, users who watch online videos for learning show different patterns compared to traditional classroom learning. For example, the high dropout rate of learners is known to be a chronic issue of online learning [147], and learners are known to be easily distracted in online learning [179]. Plus, online learning is often unidirectional, making it more difficult for learners to interact with instructors [122]. A number of research studies have investigated how users learn through online videos to better understand their behavior [142, 131, 109, 148]. Based on the understanding of learners and video formats, these studies provided insightful design implications into how video and video platforms could be designed. For instance, based on the watching pattern of selectively watching some parts of the video, Kim et al. suggested summarizing highlights of the video [142]. Similarly, Li et al. analyzed video interaction patterns and suggested design implications for utilizing video interaction patterns to improve the learning experience [148]. Yang et al. recently introduced a video-watching interface that provides learners with estimated difficulty and relevant parts of the video that are extracted from analyzing the collective interaction logs [181].

However, these studies are limited as they investigated typical users of the platform, while underrepresented user groups, such as older adults, comprise only a small portion of the platform users [112]. To increase the accessibility of online videos, understanding how various types of learners learn using online videos is necessary. Previous work also highlighted the importance of customizing the design of video platforms for underrepresented user groups, such as visually or hearing impaired users [149, 182, 134], to improve the accessibility of video. Therefore, we investigate how older adults learn through online videos and also provide design guidelines for improving the accessibility of video learning for older adults.

2.1.2 Older Adults in Learning

Lifelong learning denotes learning happening throughout one's life [146]. However, it not only stresses the characteristic of 'lifelongness' (i.e., happening throughout one's life), but also 'lifewideness', covering learning in institutions, families, communities, and workplaces [100]. In fact, for older adults, informal learning — learning occurring outside institutions or through systematic activities — is a more prevalent form of learning than formal or non-formal learning [175].

Since lifelong learning gives older adults the opportunity to learn rapidly evolving knowledge, it is known to increase self-efficacy and keep them connected to society [96]. Plus, considering that many of them are retired or are about to retire, further learning may benefit them with additional chances of extending their career [120]. Furthermore, participation in learning can also promote life satisfaction for older adults [119]. For these reasons, lifelong learning is known to increase the wellness of older adults.

However, it is known that participation in learning decreases as age increases [151]. This could be because older adults often face physical, financial, and cognitive difficulties in pursuing lifelong learning [167, 145, 125, 129, 161]. Online learning is a potential alternative to address physical and financial difficulties: (1) online learning does not require learners to be on-site, (2) flexible time choices are available without time constraints, and (3) since low-cost instructional materials are widely available online, it may reduce financial burden [99]. As such, online learning is an attractive channel of lifelong learning for older adults.

In order to help older adults fully utilize online learning platforms, it is important to design platforms that are suitable for use by this group [180, 97, 130]. Since aging involves biological, psychological, and social changes in individuals [123], older adults' behaviors and attitudes toward online learning may be different from those of non-older adults. For example, research suggests that online learning based on the Modality Principle from Cognitive Theory of Multimedia Learning [154] — instructions should not overload the learner by using only one pathway such as visual channel — is more effective for older adults than non-older adults [173]. Furthermore, older adults' motivation toward MOOC learning differed from non-older adults; older adults' motivation to learn included improving cognition and seeking fun [180]. On the other hand, their interest level in certain topics are also different; they have a higher interest in health-related topics [152, 164]. Moreover, research has found that there exist various accessibility barriers for older adults to learn online [106, 166, 94, 159, 127, 101], such as having difficulty moving to the next lesson [106].

Although these studies aimed to explore how older adults learn, they are limited to certain aspects of behaviors or difficulties (e.g., motivation, accessibility issues), which may be insufficient to fully understand *what, why, and how* older adults are learning specifically using online videos. To this end, we aim to comprehensively understand how older adults learn through online videos by focusing on the following three points with both large-scale log data and in-depth interview sessions: (1) motivation, (2) video interaction patterns, and (3) difficulties.

2.2 Usage pattern difference in large language models

2.2.1 AI-assisted Writing Patterns

Humans have used AIs for writing assistance. While traditional AI-assisted writing mostly focused on getting assistance in editing and polishing human-written texts [40, 10], the advent of LLMs has expanded AI support to earlier and more diverse phases of the writing process [22, 12, 5]. LLMs now assist with prewriting tasks, such as brainstorming and topic selection [39, 43, 24]. Moreover, many users rely on LLMs to generate initial drafts [25].

Users can interact with LLMs freely in multiple languages [47, 19]. While it is natural to query LLMs and receive responses in one's first language, there also exist cases where users interact with LLMs using non-native languages, such as scientific writing [42, 14] and business emails [37]. Non-native speakers (NNSs) exhibit distinct writing patterns compared to native speakers (NSs) due to differences in language proficiency [18, 17], which can affect clarity, accuracy, and linguistic diversity [8, 26]. For example, NNESs heavily rely on paraphrasing features in AI tools [16], and EFL students frequently use Google Translate as a dictionary and translation of challenging words [32]. Understanding these variations can help the design of LLMs, enabling them to better accommodate diverse interaction patterns and offer more personalized, effective support.

Users also exhibit varying interaction patterns with LLMs depending on the stakes involved [25]. When stakes are low, users may overlook issues with the writing experience, rely heavily on LLM-generated content, and fail to explore ways to optimize collaborative writing outcomes. This contrasts with higher-stake situations, where users are generally more engaged in refining and customizing LLM outputs to achieve better quality and alignment with their intentions.

2.2.2 User Challenges while using LLM

There are several challenges that users may encounter when interacting with LLMs. Prior works have highlighted several problems that non-native speakers (NNSs) face when interacting with LLMs. One significant issue is prompt design, as NNSs often struggle to formulate effective prompts due to limited language proficiency [35], making it difficult for them to articulate their intent accurately [6]. Many studies have also highlighted the inherent bias of current LLMs towards English, such as generating higher quality responses and comprehending prompts better [20, 26, 19], presenting an additional barrier for NNSs to achieve their desired outputs. Moreover, LLMs often misinterpret prompts written by NNSs, resulting in undesirable behaviors, such as generating less accurate or even misinformative responses [34]. Furthermore, many NNSs struggle to understand and interpret LLM responses [28, 13].

Another crucial factor for user challenges is domain knowledge. Many users lack the specific domain knowledge needed to write queries effectively, such as in medical [44] and legal domains [33]. Without sufficient understanding, users may miss key contextual details, which can lead LLMs to generate generic or incorrect responses [29]. To mitigate this issue, researchers have developed interfaces that support non-experts in formulating prompts that are similar to experts [27].

Also, familiarity with LLMs plays a key role in the user's ability to utilize LLMs, such as crafting effective prompts. Experienced users, who better understand the model's capabilities and limitations, can adjust their prompts to achieve more precise outputs. In contrast, less experienced users often struggle with prompt engineering, relying on intuitive but less systematic approaches that may result in ambiguous or incomplete responses [45].

Lastly, handling ambiguity in natural language interactions remains challenging. User queries are often ambiguous [31, 23], making it essential to resolve this ambiguity by either presenting all possible answers [3, 15] or by asking clarification questions [46, 21]. Additionally, LLM responses themselves may be ambiguous or vague, which can further complicate user understanding and lead to misinterpretations.

2.3 Perception difference in behavior log-based personality detection systems

A rich body of previous work has focused on understanding users' perception such as privacy concerns, unwanted change in behavior, and trust in results towards machine learning systems utilizing their data. Many researchers have studied users' privacy concerns towards data collection, as privacy concerns affect one's mental wellbeing, productivity, and creativity [51]. For instance, users' acceptability of sharing data significantly varies between data collected from a public and private space [74]. Users may even try to avert sharing data by using backchannels with alternative instant messaging apps or social media when they had to share even personal chats from messengers or social media [49]. Further, connotations linked to data may affect willingness to share the data: users prefer sharing information with positive connotations (e.g., step counts) than negative connotations (e.g., stress levels) [87].

Another stream of work focuses on unwanted behavior change during behavioral data collection. Oftentimes, behavioral assessments are obtrusive, i.e., users become aware of the observation, which can induce reactivity, thus changing users' natural behaviors that are significantly different from their natural behaviors [72]. Behavioral data collection is not an exception; using accelerometers to measure physical activity can also cause unwanted behavior change, increasing the first few days' amount of activity [62]. Foley et al. [63] found reactivity with a pedometer as a result of providing feedback on their physical activity.

Users' trust in machine-generated decisions or information has been an active research area. Perceived accuracy in a machine learning model can be different from the real accuracy: research reveals that humans do not trust systems of which they witnessed the mistakes, despite their high accuracy, thereby causing *algorithm aversion* [61]. On the other hand, recent research suggests that users trust algorithms over humans, i.e., *algorithm appreciation*, regardless of the domain or age [77]. Yin et al. [243] found that laypeople's trust in ML models is affected by both the model's stated accuracy and its observed accuracy in practice. This highlights the importance of understanding users' perception towards ML models.

While there has been active research on automatic personality detection in recent years [54, 90], few studies have attempted to understand users' perspectives towards APA systems. Gou et al. [67] have investigated how various factors including users' own personality and perceived benefits and risks influence users' sharing preferences of derived personality traits. In addition, Warshaw et al. [89] note that users found the automatically-generated text describing their personalities creepily accurate, but would not like to share it. Likewise, previous work on understanding users of APA systems focuses on the detected personality result, rather than on how the design of APA systems can affect users' perception. Therefore, we attempt to contribute a deeper understanding of users' perception towards APA systems across various dimensions.

2.4 Perception difference in automatic speech recognition systems

2.4.1 Bias in AI Systems

Previous research demonstrates that AI systems and their algorithms can maintain societal prejudices due to their skewed or imbalanced training data [218]. Since AI systems are dependent on the observable characteristics of the data (e.g., gender, age, skin color, regional accents), they will learn and reflect certain biases in their outputs. Buolamwini et al. reported that facial recognition algorithms most frequently misclassify darker-skinned women whereas lighter-skinned men show the lowest error rate [193]. Moreover, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software used by the US courts was found to be negatively biased against African-Americans regarding the risk of re-offending, resulting in an unfair decision-making process like harsher sentences [227].

Similarly, ASR systems are found to have certain biases as well. ASR systems are typically trained with English speech data usually spoken by people with certain characteristics, underrepresenting diverse styles of speech [214]. Feng et al. [203] and Koenecke et al. [211] found that ASR systems struggle with speech with large variations—such as gender, age, speech impairment, race, and accents—due to the training set with limited speech diversity. Tatman et al. also found that ASR systems show different performance—word error rate (WER)—between dialects and races; white speakers using Mainstream US English showed the lowest average error rate [232]. Current ASR systems are struggling to achieve high accuracy across diverse speaker groups.

Despite these known biases in ASR systems, little prior work has examined how the users of the ASR systems, i.e., the listeners, would perceive the ASR results even when the ASR systems achieve similar performance between NS and NNS. If the listeners perceive differently despite the same performance level, this may also cause biased and unfair situations. Therefore, our work investigates how listeners perceive the ASR results differently according to whether the speaker is NS or NNS.

2.4.2 Users' Perceptions of ASR System

Previous research has found that the perceived accuracy may significantly differ from the calculated accuracy of AI systems. Papenmeier et al. found that even with classifiers with equal accuracy, the prediction mistakes made on impossible-to-classify sentences were perceived to be significantly more accurate than easy or difficult-to-classify sentences [225]. Yin et al. also found that people's trust in the model can be significantly affected by the perceived accuracy regardless of its calculated accuracy as well [243].

This can be applied to the ASR systems as well. The performance of the ASR systems is generally calculated with word error rate (WER). However, Mishra et al. [217] found that while WER considers all words as equally important, in practice, users' perceptions of errors are not the same. They noticed that the users' satisfaction regarding the ASR result is dependent on the severity of the errors (e.g., incorrect recognition of names or phone numbers compared to deletion/substitution of function words like *a* or *the*) than just on the overall performance. This is because listeners judge and perceive the accuracy based on whether the transcription of the ASR systems captures the meaning of the speech.

However, little research has focused on how the users' perceived accuracy of the ASR can be different even with the same ASR result depending on whether the speaker is NS or NNS. Hence, our study focuses on how the perceived accuracy of the same caption with the same performance differs depending on NS or NNS.

2.4.3 Listeners' Perceptions of NNS's Speech

Previous research has investigated the negative influence of NNS's accent on listeners. Lev-Ari et al. found that the NNS's accent serves as a signal that the speaker is an out-group member as well as a factor that makes the speech more difficult to understand [213]. They showed that accented speech reduces "processing fluency". Moreover, the listeners perceived accented speech as less truthful and perceived NNS as less credible [213]. Similarly, it was found that the listeners make lower ratings in terms of attractiveness, benevolence, and trustworthiness to speakers with accents that are considered foreign or spoken by minorities) [204]. Consequently, listeners' different perceptions according to one's accent may bring severe social consequences such as discrimination [190].

Nonetheless, little prior work has investigated how listeners perceive the speaker differently according to whether the speaker is NS or NNS when they also get exposed to the ASR output of their speech. Since ASR technologies are frequently used in computer-mediated communications to support multilingual communication settings, it is important to investigate how ASR technology affects the listener's perception. Hence, our study attempts to investigate how different qualities of captions change the listener's perception of NNS's speech, as well as how this prior perception may result in blaming the speaker for the errors in the ASR-generated captions.

Chapter 3. How Older Adults Use Online Videos for Learning

Online videos are a promising medium for older adults to learn. Yet, few studies have investigated what, how, and why they learn through online videos. In this study, we investigated older adults' motivation, watching patterns, and difficulties in using online videos for learning by (1) running interviews with 13 older adults and (2) analyzing large-scale video event logs (N=41.8M) from a Korean Massive Online Open Course (MOOC) platform. Our results show that older adults (1) are motivated to learn practical topics, leading to less consumption of STEM domains than non-older adults, (2) watch videos with less interaction and watch a larger portion of a single video compared to non-older adults, and (3) face various difficulties (e.g., inconvenience arisen due to their unfamiliarity with technologies) that limit their learning through online videos. Based on the findings, we propose design guidelines for online videos and platforms targeted to support older adults' learning.

3.1 Introduction

"One is never too old to learn," "Learning is from cradle to grave." As emphasized by these proverbs, lifelong learning, which spans from early childhood to older age, is crucial to one's life. Lifelong learning not only gives one a sense of personal fulfillment and satisfaction [116, 124, 107], but also enables them to adapt to a fast-evolving job market [107]. Furthermore, it strengthens a nation's economy and prevents exclusion or marginalization of older adults [118, 124].

Among various media that could support lifelong learning, online videos are among the most popular due to their availability, scalability, and cost-effectiveness [157]. For example, video-based learning platforms, such as Massive Online Open Course (MOOC) platforms, are widely available these days. Additionally, video platforms like YouTube are offering instructional videos such as how-to videos in diverse domains (e.g., cooking, swimming, fishing).

With the rise of online video learning, a myriad of research has investigated how users use online videos for learning, which has provided insights on how to design videos and tools that further enhance the learning experience [103, 160, 142, 131]. However, older adults who are retired or in the later stages of their careers may exhibit different usage behaviors due to possible age-related factors (e.g., sensory and perceptual issues, slower processing speed, low working memory abilities) [123]. Moreover, in contrast to the younger "video generation", older adults are likely to be accustomed to one-way interaction with video (e.g., TV) [144, 168]. Thus, their use of online learning videos might be different from that of non-older adults. However, little research has examined *what, why, and how* older adults use online learning videos to learn. Understanding how older adults learn through online videos would be crucial to providing an appropriate aid for such a segment of the population to facilitate the use of online videos for learning.

To this end, in this paper, we aim to understand how older adults (i.e., those aged 55 or older[1]) use online learning videos in terms of (1) what videos they watch online for learning and why, (2) how they interact with online videos, and (3) what difficulties they face. We investigate these aspects through a mixed-methods approach: (1) interviews with 13 older adults ($M_{age} = 65.5$, $SD_{age} = 6.6$)

[1] These ages are indicated in Korean age. Korean age considers the birth year as year 1, which is equivalent to calculating the age as $currentyear - birthyear + 1$.

who have used online videos for learning and (2) large-scale log analysis of older adults' interaction logs on a MOOC platform in comparison with those of non-older adults (41.8M interaction logs from total 108K users). We found that older adults tend to watch online videos to learn practical topics applicable to their daily lives (e.g., English conversation, cooking), while consuming fewer videos in science or engineering domains. Moreover, we identified that they (1) perform fewer video interactions (i.e., pause, jump forward/backward), (2) watch videos more repeatedly, and (3) cover a video[2] more than non-older adults. We also identified that older adults face difficulties due to (1) the characteristics of the video medium and technology and (2) video-specific issues (e.g., fast speaking pace). Based on the findings, we suggest design implications of online videos and their platforms for older adults to better pursue lifelong learning.

The contributions of this paper are as follows:

- Results from an analysis of 41.8M log events and interviews with older adults that reveal why and how older adults watch online videos for learning and the difficulties older adults face when watching online videos for learning
- Design guidelines of online videos and their platforms for older adults to have a better learning experience using online videos

3.2 Method

We took a mixed-methods approach, incorporating both interviews and a large-scale MOOC log analysis, to understand how older adults use online videos to learn. By analyzing older adults' video usage logs, it is possible to understand how older adults watch videos for learning from their natural behavior logs and how their behaviors differ from those of non-older adults. However, understanding why they show such behaviors and what difficulties they face might be limited with the log analysis alone. Thus, we conducted interviews with older adults in addition. We used an emergent mixed-methods design [113], where we first started with log analysis only and later conducted interviews to draw complementary insights. Note that the logs analyzed were collected in 2018 and the interviews were conducted in 2020.

Our research questions are as follows:

- **RQ1**: [**Motivation**] Why and what do older adults want to learn while watching online videos?
- **RQ2**: [**Watching pattern**] How do older adults watch online videos for learning?
  - **RQ2-1**: What do they consider when choosing which videos to watch?
  - **RQ2-2**: How much do they interact with the video?
  - **RQ2-3**: How and why do they watch a video repeatedly?
  - **RQ2-4**: How much of the video do they watch?
- **RQ3**: [**Difficulties**] What are the difficulties older adults face while learning through online videos and how do they try to address the difficulties?

We combined the interview and log analysis to gain a comprehensive understanding. We originally started our log analysis for RQ1 and RQ2-2. Upon realizing that exclusively relying on the log analysis

[2] We define the coverage of a video as the percentage of a video clip seen by the viewer.

provides a limited understanding of older adults' video-based learning behavior, we decided to conduct interviews to complement the findings from the log analysis. While planning the interview study, we added RQ2-1 (how they select videos to watch) and RQ3 (difficulties) as these are aspects that are essential in understanding the end-to-end process of how older adults learn with videos. These were also questions that the log analysis could not answer. We also further added RQ2-3 and RQ2-4 for clarity in reporting. After the interview, we conducted an additional log analysis to answer RQ2-3 and RQ2-4, as these are the research questions that were added later that could be also answered through log analysis in addition to the interview.

Definition of Older Adults Although most previous studies defined *older adults* as those whose age spans 55 to 65 [121, 155], no fixed agreement exists on which chronological age could define older adults. This is because the term 'older adults' have different criteria based on their societal surroundings, ranging from family to culture or world [174].

As such, we refer to Findsen et al. [124] to define older adults as people who are having or are about to have a later stage of life. Specifically, since the average retirement age in Korea was around 57 in 2018 [137], when our data was collected, we defined older adults as those who are aged 55 or older in Korean age.

Scope of Online Videos for Learning The goal of our research is to investigate how older adults use online videos in general for learning purposes. Since the types of online videos one can learn from vary greatly, we interviewed those who watched *any video* if they watched the video with the *purpose of learning* to capture diverse watching experiences. While granting an in-depth account of individual experiences, interview results can only capture behaviors participants remember. Thus, for a more comprehensive understanding of the landscape, we complemented interviews with a log analysis from a large-scale video platform.

We chose a MOOC platform for the log analysis, since we can ensure users of a MOOC platform watched MOOC videos for learning purposes. Most MOOC platforms, however, have several differences from other video platforms; they usually offer videos in limited styles [136] or topics and many take courses for college credits or certificates. Insights gathered from analyzing a MOOC platform would lead to a misalignment with more generic video learning experiences of our interview participants. Considering this, we chose to analyze data from K-MOOC[3] [140] in 2018 due to its breadth of topics and video styles and credit system not being adopted yet. First, K-MOOC provides videos with various topics aiming for providing lifelong learning ranging from common MOOC topics (e.g., Using Python for Big Data Analysis[4], Reading American Literature with Pictures[5]) to various topics related to daily life (e.g., Smoking and Healthy Life[6], All about My House[7], Creative People's Seven Habits[8]). Second, format-wise, they are not only limited to typical styles of MOOC videos, but include various formats such as practicing workout steps, narrative animation [136], or fictional case study [136]. Lastly, in 2018, K-

[3] A Korean state-led MOOC platform, which launched in 2015 with 3.6M users by the end of 2018. It offered 520 courses open for enrollment as of January 2019, spanning various subject domains (e.g., humanities, social science, engineering, natural science) offered by 92 different universities.

[4] Domain: Engineering, Level: Intensive major

[5] Domain: Humanities, Level: Basic major

[6] Domain: Medical sciences & Pharmacy, Level: Elective

[7] Domain: Engineering, Level: Elective, A course that covers how to pick a good home, how to interior the house, how to invest using real estate, and knowledge for house taxation

[8] Domain: Social science, Level: Elective

Table 3.1: Participants of the interview sessions

ID	Age	Gender	Education	Domains of videos watched for learning
P1	76	M	Master	Bible, health, English conversation
P2	65	F	Master	Biblical Hebrew, Bible, theology
P3	57	M	Master	Work-related IT field, statistics, deep learning, cookery, camping
P4	56	F	High school	Sports, health, diet, preparation for old age
P5	69	F	Doctorate	Chinese, calligraphy, DIY, musical instrument, gardening, cookery, health, life wisdom
P6	63	M	Bachelor	Photoshop, camera, Chinese, astronomy, fire safety
P7	74	F	Bachelor	Farming, cookery, sports
P8	75	M	Master	Oil painting, farming
P9	65	M	Bachelor	Counseling studies, golf, yoga
P10	57	F	High school	Stock investment, storytelling (Korean traditional stories), cookery
P11	64	M	Master	English (conversation, vocabulary), golf, fishing, billiard
P12	65	F	Bachelor	Taxation, cookery, interior architecture, health, astronomy
P13	65	F	Bachelor	English, swimming, cookery

MOOC has not yet adopted the Academic Credit Bank System [93]; it did not allow one to earn credits for earning a degree. In summary, we believe K-MOOC—with its broad topical coverage and lifelong learning support—can serve as a compatible source of data to complement our interviews—with more generic video learning experiences.

3.2.1 Interview

We recruited 13 adults aged 55 or older who had experience watching online videos for learning within six months. We posted recruitment ads in online communities where older adults are expected to visit (e.g., online bulletin board targeted for 50s+), along with online communities where the users' parents may be in the age of older adults (e.g., online communities of colleges) to recommend their acquaintances who qualify. We tried to diversify the interviewee pool by considering their age, level of education, pre-reported frequency/amount of learning using videos and selected 13 interview participants (7 females, M_{age} = 65.5, *SD* = 6.6) (Table 3.1). All interview sessions were conducted through voice calls and lasted around 60-90 minutes. Each session was audio-recorded, and each participant received 25,000 KRW (22 USD equivalent) for their participation. The study was approved by our institution's IRB.

We conducted semi-structured interviews, where we asked (1) *personal information* (e.g., age, education degree), (2) *general experiences on learning through online videos* (e.g., motivation for learning, first time to start learning through online videos, how they got accustomed to online video interface), and (3) *experience on learning through online videos for each video they mentioned* (e.g., content and form of the video, motivation, how they watched the video, other activities they did relate to the video, difficulties they faced and how they resolved them). All interviews were transcribed and then analyzed with a thematic analysis [108, 172].

Two researchers conducted thematic analysis by first reading transcripts and noting notable patterns of behaviors or quotes. Then, we classified notes into the most relevant research question. For each research question, we classified notes into theme. To improve coherence within the theme, we iterated over the notes within each theme while re-classifying a subset of notes. Here, we discussed on the note categorization where we disagreed to reach a consensus. Finally, we labeled the themes. For RQ2-2 and RQ3, researchers agreed that there exists a need for classifying further into subthemes — to identify detailed reasons behind the identified interaction behavior themes (RQ2-2) and to further classify the high-level challenges to identify the detailed reasons behind the challenges older adults face (RQ3). Thus, we further decomposed the notes in each theme into subthemes by going over the same process with when identifying the themes. Finally, one researcher re-examined the notes of themes and subthemes for coherence.

3.2.2 Log Analysis

We analyzed event log data from K-MOOC collected in 2018 to understand how older adults use MOOC videos compared to non-older adults. Through comparison with typical users, we wanted to understand how older adults are *unique* in their way of using the videos, as it could provide insights into better designing current online videos and their platforms customized for older adults. Specifically, we took into account their video domain selection which indirectly reveals their motivation (RQ1), frequency of single interactions and watching patterns (i.e., interaction sequence) (RQ2-2), length of repeated watched parts (RQ2-3), and coverage per video (RQ2-4) as dependent variables, while age group (i.e., older adults and non-older adults) being the common independent variable of all the log analysis.

Data and Pre-processing

The event logs capture users' video interactions (i.e., play/stop/pause video and seek back/forward) on their interaction type, video timestamp, real-time, and user & course information, across 1.4K different courses and 51K different lecture videos (See Supplementary Material for sample logs). Data were provided from the K-MOOC platform upon the grant contract after all personally identifiable information had been anonymized. After excluding the logs with errors that are not recoverable (e.g., (1) missing certain fields describing an event (e.g., the time when an event occurred) and (2) having duplicate values for certain fields (e.g., having two different times for an event)) and extracting video-related event logs as of our purpose, 41.8M event logs were left. These video event logs included behaviors of 108K different K-MOOC users on 1,391 different courses. Among the users who provided their birth year when signing up (107K users), 4.4K users (2.8% of all users) were classified as older adults in 2018 (Figure 3.1).

Additionally, we also obtained sign-up information (i.e., birth year, gender, etc., which users optionally entered while signing up) of users who signed up until 2018 (3.6M users) and information of 438 courses that were open for enrollment in 2018, including course name and subject categorization. Since video length information was not stored in the database as a separate entry, we extracted the length of 23.7K videos from 476 different courses, which we were able to access at the time we crawled (April 2020).

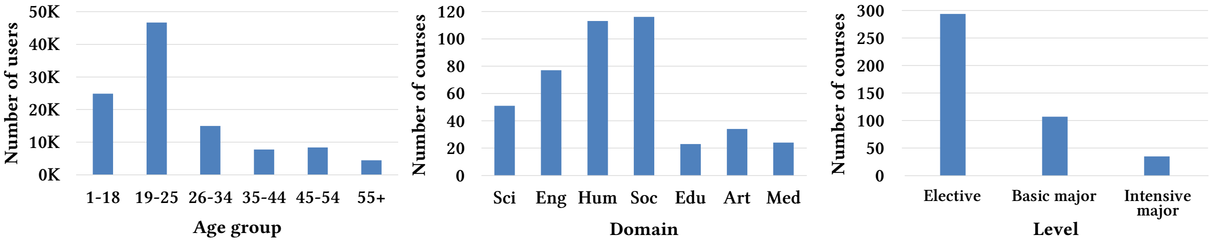


Figure 3.1: Age distribution of users who watched K-MOOC videos in 2018 (left), distribution of courses by domain (middle) and level (right) offered in K-MOOC in 2018 (Sci: natural science, Eng: engineering, Hum: humanities, Soc: social science, Edu: education, Art: arts & physical education, Med: medical sciences & pharmacy)

RQ1: What do they want to learn while watching videos?

Analyzing how older adults select courses, in which domain and level of difficulty (i.e., elective, basic major, intensive major), may give us insights into what older adults want to learn using online videos. Therefore, we analyzed how the domain selection and level selection of older adults differ from that of non-older adults. We based the categorization of each video on K-MOOC's classification, which was determined by the instructor: seven high-level domain categories and three levels of difficulty as in Figure 3.1. To avoid taking cases into account where a user may have clicked a video mistakenly, we only considered the courses the user took with at least three log events (i.e., play, stop, pause, seek, changing speed, and showing/hiding captions or transcript). Then, we used logistic regression to identify the relationship between the age group (i.e., older adults and non-older adults) and whether the user will take a course in each domain category. As there exists a correlation between each case since individuals took multiple lectures and the same lecture video is watched by multiple users, we used generalized estimating equations (GEE) model [133], a statistical method used when correlation may exist in the outcome variable. We used the exchangeable correlation structure as one may watch a course at different times watching several videos, so their order of watching courses may change.

RQ2-2: How much do they interact with the video?

Analyzing how older adults interact with the video would provide insights into how they watch a video [141]. Thus, we (1) analyzed the frequency of video interactions and (2) performed sequence clustering [176] to know the dominating interaction sequence pattern to understand how each older adult watched each video. Among various types of video interactions, we particularly focus on watch, pause, and seek forward/backward[9] for the analysis, as they have a direct connection with the flow of how users consume video content, unlike speed change or turning on/off captions or subtitles.

1. Frequency of single interactions. Even for the same jump, the intention behind performing a long jump could be different from a short jump. Thus, we subclassified each interaction into three detailed interactions based on the length or duration of the interaction. As in Table 3.2, the threshold between short and medium interactions is determined as 25 percentile of interaction length/duration, while the threshold between medium and long interactions is determined as 75 percentile of interaction length/duration. For 'watch' interaction, unlike other interactions like pause or seek where pause/seek

[9] In this paper, we define 'seek backward' as jumping to a prior part of the video and 'seek forward' as jumping to a later part of the video

Table 3.2: We defined 12 detailed interactions, based on the length or duration of each interaction.

Type	Detailed name of interaction	Definition
Watch	Short Watch (SW)	0.2 s $\leq$ Watched duration of video timestamp $<$ 1.8 s
	Medium Watch (MW)	1.8 s $\leq$ Watched duration of video timestamp $<$ 44.2 s
	Long Watch (LW)	44.2 s $\leq$ Watched duration of video timestamp
Pause	Short Pause (SP)	Paused duration $<$ 1.6 s
	Medium Pause (MP)	1.6 s $\leq$ Paused duration $<$ 45.0 s
	Long Pause (LP)	45.0 s $\leq$ Paused duration
Seek Backward	Short Seek Backward (SB)	-8.2 s $\leq$ Seek videotime length $\leq$ 0s
	Medium Seek Backward (MB)	-33.6 s $\leq$ Seek videotime length $<$ -8.2 s
	Long Seek Backward (LB)	Seek videotime length $<$ -33.6 s
Seek Forward	Short Seek Forward (SF)	0s $<$ Seek videotime length $<$ 9.0 s
	Medium Seek Forward (MF)	9.0s $\leq$ Seek videotime length $<$ 36.6 s
	Long Seek Forward (LF)	36.6 s $\leq$ Seek videotime length

interval begins with a user pressing the ‘pause/seek’ button, a watched interval can begin without the user actually pressing the ‘play’ button at the start of the interval. Thus, we decided to ignore the cases where only watching status lasts less than 0.2 seconds. Moreover, as we are defining detailed interactions according to their length relative to the video length while the length of the videos in K-MOOC varied a lot (m = 13.5 minutes, std = 9.4 minutes), we focused only on the logs that were performed in the videos that have the length that falls into the 25 to 75 percentile of the video length distribution: 6.95 minutes to 18.07 minutes.

Then we calculated the frequency of each detailed interaction each user performs in each video. We took the following two metrics to calculate frequency to capture complementary aspects: (1) Frequency 1: number of times each detailed interaction is performed per minute; calculated by dividing the number of times each detailed interaction is performed by the total length of the corresponding video, and (2) Frequency 2: number of times each detailed interaction is performed per coverage of the video one watched; calculated by dividing the number of times each detailed interaction is performed by the coverage of the corresponding video that the learner watched at least once. For Frequency 2, we excluded cases where one’s coverage of the video is less than 1% to avoid dividing by near zero.

We then used linear regression to identify the relationship between the frequency of each detailed interaction and the user’s age group. This is because through RQ1 we found that older adults and non-older adults watch videos of different domains and levels of the video, where the domain and level of the video may affect the frequency of interaction. Thus, we used regression while taking the domain and level of the video as regressors in the model to understand the relationship between frequency of interaction and the user’s age group without these factors affecting the result. Similar to the reason explained in RQ1, we used Generalized Estimating Equations (GEE) with an exchangeable correlation structure. In addition, in order to compare which detailed interaction has more frequency difference between older adults and non-older adults, we standardized the dependent variable and iterated upon the same condition.

2. Dominating interaction sequence pattern. Although we can know how frequently older adults exhibit different interactions by analyzing individual interactions, it does not capture patterns of video watching at a macro level. There may be cases where a sequence of interactions signals a specific intent. For instance, even with the same number of Short Seek Forwards performed within a video, one could be seeking forward to look for a specific part or to skim the whole video. Thus, we analyzed the interaction *sequence* by sequence clustering method proposed by Wang et al. [176] to identify older adults' emergent video-watching patterns and how they differ from those of non-older adults.

We first converted each learner's interaction logs of a video in a session to a watching pattern sequence composed of interaction units defined in Table 3.2. Then, we extracted every possible subsequence of length k (i.e., k-gram sequence) from the watching pattern sequence. For every two watching pattern sequences pair, we calculated the normalized frequency per subsequence that appeared in either of the two sequences. These normalized frequencies are made into an array per sequence. Next, we used polar distance between the two arrays to cluster the sequences. (Refer to Wang et al. [176] for more detailed information on sequence clustering.) We chose k as 4 in k-gram sequence as repetitiveness is captured enough in 4-gram (See Supplementary Material for details). Furthermore, similarly to analyzing the frequency of each detailed interaction, we focused only on the videos that belong to the 25 to 75 percentile of video length distribution. Then, due to time complexity, we randomly sampled a total of 20K watching sequences of a user watching a video in a session (i.e., 10K from older adults and 10K from non-older adults) for sequence clustering. With the sequence clustering results, we made dummy variables for each cluster and ran GEE with a binary exchangeable correlation structure.

RQ2-3: How and why do they watch a video repeatedly?

To understand how older adults watch videos repeatedly, we extracted the sum of the lengths of all the repeatedly watched parts. If a user watched a part three or more times, each repeated watching time was added. As the length of videos varies, we calculated the percentage of the length of all the repeatedly watched parts by dividing by the length of the video. Next, we used GEE model with linear regression to identify the relationship between the percentage of the repeated watch and the age group. For a similar reason to the previous analyses (RQ2-2), we also considered the domain and level of the video as factors in the model and used an exchangeable correlation structure. We also took the length of videos as a factor in the model as we did not limit the analysis to a certain length of the videos.

RQ2-4: How much of the video do they watch?

To identify how much older adults cover a video, we extracted the sum of lengths of all the watched parts, regardless of the number of times watched. Then, we derived the coverage of the video by dividing it by the length of the video. Next, we used GEE model with linear regression to identify the relationship between the coverage of a video and the user's age group in the same setting as with RQ2-3.

3.3 Results

We present the results of the thematic analysis of the interviews and log analysis for each RQ. For the thematic analysis result, we present the identified themes of all RQs. For the log analysis, we present the results of RQ1, RQ2-2, RQ2-3, and RQ2-4.

3.3.1 RQ1: Why and what do they want to learn while watching online videos for learning?

Interview

We identified two themes on what older adults want to learn — (1) those related to their personal interests, hobby, curiosity, or needs in their daily life and (2) those related to their work.

We found that *all participants wanted to learn at least one* subject **related to their personal interests, hobbies, or needs in their daily life**. This relates to more self-directed or autonomous learning rather than required learning [98]. What they learned spanned across various disciplines, including visual arts, physical education, humanities, social sciences, and even practical life skills (Table 3.1). In contrast, only three participants said they had learned a STEM subject. Some participants attributed the lack of desire to learn STEM to the difficulty of learning: *"For me, it's hard (to learn scientific topics)"* (P4). P7 even explicitly mentioned that they regard learning something completely new as not suitable for their age. Even among the participants who watched STEM videos, it was largely limited to the surface level (2 out of 3 participants). For example, P12, who watched astronomy videos, said: *"Although I have interest in science, I'm not interested in the theories but watch (science videos) for their awe-inspiring feelings" (P12).*

Four participants mentioned that they watch online videos to learn things **related to their work**, although no participant reported watching online videos solely for learning work-related materials. This is linked with required or mandated learning [98]. In this case, they were more driven by external factors, including learning something that relates to their job (e.g., speaker system development, health education, deep learning basics).

Log Analysis

Results show that older adults take more courses in humanity and medical science while taking fewer courses in STEM, social science, and education, compared to non-older adults. This aligns with our interview result that not many older adults watch STEM domain videos. Furthermore, it aligns with previous research [139, 152] that older adults tend to like learning about health science. The *odds ratio* [105] of each domain, which indicates the ratio of the odds of older adults taking a course in a certain domain to the odds of that of non-older adults, is presented in Figure 3.2. For example, the odds of older adults taking natural science courses are 0.65 times that of non-older adults. Except for arts & physical education courses, there exist clear differences between the odds of older adults taking a course in a certain domain compared to those of non-older adults.

Moreover, compared to non-older adults, older adults prefer taking elective courses over major courses. The odds ratio decreases as the level increases (elective courses: 1.14, basic major courses: 0.90, intensive major courses: 0.70) (Figure 3.2). This may be the result of older adults trying to learn something related to their personal interest or curiosity rather than for their work, as shown in our interview result.

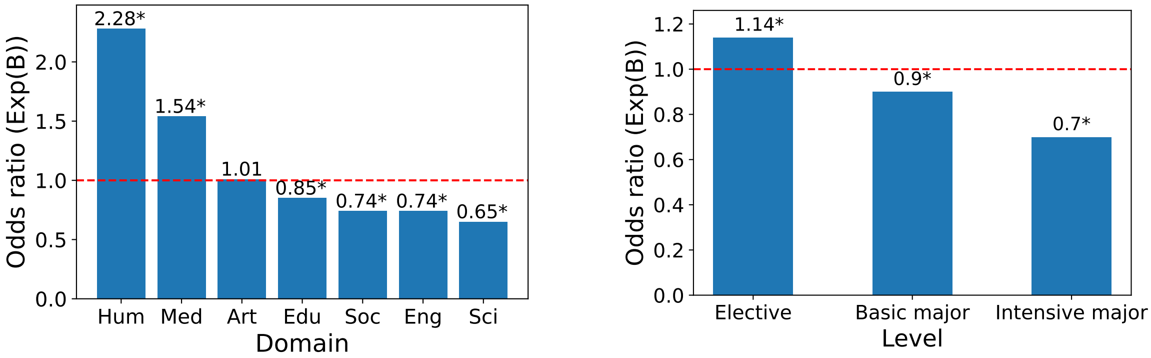


Figure 3.2: Results of RQ1, which displays the odds ratio for the domain (i.e., Humanity, Medical sciences & Pharmacy, Arts & Physical education, Education, Social science, Engineering, Natural science) and level (i.e., Elective course, Basic major course, Intensive major course). Older adults take more humanity or medical sciences & pharmacy courses and fewer engineering, natural science, social science, and education courses than non-older adults. They also take more elective courses and major-related courses than non-older adults. (* indicates *p ¡ .01*)

3.3.2 RQ2-1: What do they consider when choosing which videos to watch?

Interview

The criteria for deciding what to watch have emerged as follows: (1) video metadata, (2) whether the content and level suit their expectations, and (3) whether the video is in their desired format.

Nine participants mentioned **video's metadata** (e.g., title, thumbnail, uploaded date, creator/uploader) as the main criteria for selecting the video to watch: *"If I have something I want to know, I first search their (i.e., prominent tax accountant's) channel . . . I don't think his delivery is better, but I can trust what he's saying"* (P12). However, some mentioned that deciding with metadata would lead to misselection. Thus, some mentioned that they would only continue watching when it fits the additional criteria (e.g., content, level, format).

Ten participants also mentioned **desired content and difficulty level of the video** as criteria for choosing a video to watch. Specifically, they wanted to find a video that covers the contents that fit their level. According to P9, since the metadata does not mention whether the yoga video is for older adults, they should watch the video to judge whether it suits their level and stop watching if it does not. Interestingly, four participants pointed out that the creator being in their age is an indicator of proper content and level: *"I don't really understand when watching cooking videos made by young people. Moreover, while I prefer cooking Korean dishes, they usually cook Western dishes."* (P10).

Twelve participants pointed out the **desired format** as a criterion for selecting videos to watch. Their desired format includes demonstration with appropriate close-ups to how-to videos, having a feedback session, and showing more visual materials than just explaining. Interestingly, many (8 participants) favored a format that delivers the core content without adding jokes, irrelevant chats, or advertisements: *"I don't think trying to be funny or interesting is necessary. (I like it when they tell me) just the key points"* (P7).

3.3.3 RQ2-2: How much do they interact with the video?

Interview

We identified emerging themes of (1) not performing many interactions overall and (2) performing seek forward, seek backward, or pause time to time, where the former took most proportion. Thus, we report the subthemes that represent the reasons behind the first emergent theme.

When asked to describe how they interact (e.g., pause, seek forward) with the video while watching it, eight participants replied that they **do not really perform many interactions overall**. This was in part because they did not find it necessary as they could understand the video content but also because they did not need to fully understand the content, which is linked to their motivation of watching the video (RQ1). Moreover, they did not interact with the video because they did not *know how* to interact with the video or felt uncomfortable interacting with the video. P4 mentioned, *"I didn't know about pausing or other functionalities . . . So I watch (the video) from the beginning again."*

Particularly for the seek forward interaction, six participants said they do **not mostly seek forward at all**. The most prominent reason was that they prefer not to miss or skip anything. P1 said: *"(Even if I watch a video several times, if I seek forward, I feel like I'm learning less even though it may save some time."* Similarly, P9 mentioned: *"(Although I'm only looking for a certain part in a video,) I watch it from the beginning without seeking forward. It is because I want to know everything . . . (It's also because, even if I am knowledgeable about the other part of the contents,) explanation styles across lecturers may vary"*.

Log Analysis.

We report results from analyzing (1) the frequency of video interactions and (2) dominant interaction sequence patterns.

1. Frequency of single interactions. We found that older adults perform significantly fewer interactions while watching online videos for learning compared to non-older adults (Table 3.3, Table 3.4). Moreover, the negative values of all coefficients in Table 3.3 and Table 3.4 indicate that older adults perform all detailed interactions less than non-older adults. This aligns with our interview result of older adults not performing interactions a lot.

To identify which interaction has a larger frequency difference between older adults and non-older adults, we standardized the dependent variable and ran the model again, whose result is shown in the last column of Table 3.3 and Table 3.4. Moreover, older adults tended to perform large seek forwards much less compared to non-older adults. This may be the result of not performing seek forward as they do not want to miss anything as seen in the interview. They also exhibit short watches much less than long watches compared to non-older adults. As the watched interval was defined by watched segment without pause or seek interaction, this also strengthens the interview result of older adults not performing interactions overall.

2. Dominating interaction sequence patterns. We identified seven sequence clusters along with the top three sequence patterns that are prevalent in each cluster, distinguishing the cluster from other clusters (Table 3.5). We also grouped sequences that were not included among the seven clusters as Cluster *Etc.*. The odds ratio of each cluster in Table 3.5 indicates the ratio of the odds of older adults watching a video in the pattern of the corresponding cluster to the odds of that of non-older adults. For

Table 3.3: Results of RQ2-2-1: distribution and the result of linear model regression using GEE for Frequency 1 (i.e., number of times each detailed interaction is performed per minute)

Detailed Interaction	**Avg. & Std. for all**	**Avg. & Std. for older adults**	**Avg. & Std. for non-older adults**	**B (Coefficient of age group being 55+)**	**B (Coefficient of age group being 55+ after standardizing dependent variable)**
SW	0.17 / 0.51	0.05 / 0.19	0.18 / 0.53	-0.120 *	-0.237 *
MW	0.29 / 0.63	0.24 / 0.66	0.30 / 0.63	-0.069 *	-0.108 *
LW	0.13 / 0.15	0.13 / 0.12	0.13 / 0.15	-0.002 *	-0.013 *
SP	0.04 / 0.36	0.02 / 0.09	0.05 / 0.38	-0.022 *	-0.061 *
MP	0.12 / 0.39	0.10 / 0.40	0.12 / 0.39	-0.032 *	-0.082 *
LP	0.05 / 0.10	0.05 / 0.09	0.05 / 0.10	-0.003 *	-0.031 *
SB	0.04 / 0.17	0.02 / 0.10	0.04 / 0.18	-0.018 *	-0.107 *
MB	0.08 / 0.22	0.05 / 0.16	0.08 / 0.22	-0.030 *	-0.139 *
LB	0.04 / 0.11	0.03 / 0.08	0.04 / 0.11	-0.012 *	-0.109 *
SF	0.10 / 0.48	0.07 / 0.37	0.10 / 0.49	-0.064 *	-0.053 *
MF	0.10 / 0.53	0.09 / 0.34	0.20 / 0.54	-0.098 *	-0.177 *
LF	0.09 / 0.19	0.05 / 0.14	0.09 / 0.19	-0.042 *	-0.225 *

Table 3.4: Results of RQ2-2-1: distribution and the result of linear model regression using GEE for Frequency 2 (i.e., number of times each detailed interaction is performed per coverage of the video one watched). For example, from this table and Table 3.3, we can infer that the coefficient of the age group being 55+ for Short Watch (SW) is -0.120, which means compared to non-older adults, the number of times Short Watch (SW) performed per minute by older adults is on average 0.120 times/minute smaller, while the average of whole users being 0.17 times/minute. Moreover, the number of times Short Watch (SW) is performed per coverage of the video one watched by older adults is on average 0.162 times/covered minute smaller, while the average of whole users is 0.2 times/covered minute (* indicates $p < 0.01$). (Note that some averages of detailed interaction appear to be similar between older adults and non-older adults, but it could be due to the fact that each group (older adults and non-older adults) has a different distribution of watching videos in terms of levels and domains, while these factors also significantly affect the interaction frequency.)

Detailed Interaction	**Avg. & Std. for all**	**Avg. & Std. for older adults**	**Avg. & Std. for non-older adults**	**B (Coefficient of age group being 55+)**	**B (Coefficient of age group being 55+ after standardizing dependent variable)**
SW	0.20 / 0.79	0.04 / 0.27	0.22 / 0.82	-0.162 *	-0.205 *
MW	0.20 / 0.40	0.11 / 0.30	0.21 / 0.41	-0.097 *	-0.243 *
LW	0.03 / 0.03	0.03 / 0.03	0.03 / 0.03	-0.001 *	-0.038 *
SP	0.03 / 0.24	0.01 / 0.07	0.03 / 0.25	-0.016 *	-0.067 *
MP	0.04 / 0.12	0.03 / 0.11	0.04 / 0.12	-0.016 *	-0.131 *
LP	0.01 / 0.05	0.01 / 0.04	0.01 / 0.05	-0.003 *	-0.057
SB	0.02 / 0.18	0.01 / 0.09	0.02 / 0.19	-0.012 *	-0.066 *
MB	0.04 / 0.23	0.02 / 0.11	0.05 / 0.24	-0.026 *	-0.113 *
LB	0.03 / 0.18	0.02 / 0.10	0.04 / 0.19	-0.019 *	-0.099 *
SF	0.07 / 0.98	0.03 / 0.26	0.07 / 1.02	-0.041 *	-0.042 *
MF	0.21 / 1.03	0.06 / 0.52	0.22 / 1.07	-0.148 *	-0.140 *
LF	0.17 / 0.66	0.07 / 0.39	0.18 / 0.68	-0.105 *	-0.161 *

example, the odds of older adults watching with the dominating pattern of consistent medium or long seek forwards (i.e., MF & LF) without watching (i.e., MF-MF-MF-MF or LF-MF-MF-MF or MF-MF-MF-LF) are 0.634 times that of non-older adults.

We found that older adults watch in a different watching sequence compared to non-older adults; the odds ratio was significantly different except for Cluster 4. Among those, only Cluster 5 (i.e., repeated long-term watching and long-term pause), had higher odds of older adults watching in that pattern than non-older adults. This indicates that older adults are more likely to watch at a longer pace just pausing for a long time once in a while. Results also show that older adults are less likely to watch in a constant skipping or skimming manner (Cluster 2, 3, 6). This also strengthens our interview result that older adults do not prefer missing anything in addition to single interaction analysis (Section 4.3.2.1). Moreover, the odds of older adults watching in a pattern that is not common enough so that it does not belong to any clusters were around two times higher than that of non-older adults (Cluster Etc.). This indicates that they are more likely to watch in sequences that are not frequently watched by others.

3.3.4 RQ2-3: How and why do they watch a video repeatedly?

Interview

A lot of participants (11 participants) reported that they watch videos repeatedly, where three themes emerged as reasons behind rewatching: (1) to follow the videos, (2) to remind themselves of the contents, and (3) to learn and understand the contents more thoroughly.

First, participants watched videos **to follow the videos, while most of them repeatedly watched video before starting to follow** the actions in the video. Interestingly, among the participants who wanted to follow the video, six participants reported that they did not follow the video while simultaneously watching it. Instead, they preferred to rewatch the videos repeatedly until they could ultimately follow the video without watching it. They also reported that they also stick to that video for a long time and follow the video repeatedly, before shifting to another video. P9 said, *"I watch videos repeatedly until I can do the workout completely by myself without watching . . . There are only one or two videos that I have completely understood. For one video, I even watched about 20 times."*.

Second, 11 participants reported that they rewatch videos **to remind themselves of the contents**. Among them, eight participants reported that they rewatch when they forgot some contents, while others reported that they watch repeatedly since they worry they would forget the contents later: *"I usually watch around 3 to 6 times. Now my memory got worse (than when I was young). Although I think I'm better than others my age."* (P1).

Lastly, seven participants added that they would rewatch **to fully learn and understand the contents more thoroughly**. They considered that rewatching a video repeatedly is crucial to learning. P11 said, *"I'd download the video and rewatch, as (even after watching) it's not fully mine."* P6 also said, *"Even though I try to watch all the details, I can't understand everything by only watching once . . . By watching again after some time has passed, I can notice something that I haven't noticed before."*

Log Analysis

Overall, neither older nor non-older adults showed rewatching patterns frequently. The majority of learners watched videos (93.1%) with rewatching happening in less than one-tenth of the video length, indicating that the vast majority just rewatch small parts of the video. Moreover, only 0.4% of the cases rewatched more than 100% of the video, implying that watching a video multiple times was rare.

Table 3.5: Results of RQ2-2-2: Sequence clustering result and the odds ratio of older adults for each cluster. Per each cluster, the top three sequence patterns that are prevalent in each cluster, which distinguish the cluster from other clusters are presented. Percentile refers to how common the cluster is.

Cluster # **(Percentile)**	**Patterns**	**Pattern Explanation**	**Exp(B)** **($p < 0.01$)**
Cluster 1 (13.6%)	MP-MW-MP-MW MW-MP-MW-MP MP-MW-LP-LW	medium-lengthed watching with intermittent medium pause	0.758
Cluster 2 (11.4%)	MW-MF-MW-MF MF-MW-MF-MW MW-MF-SW-MF	medium-lengthed watching with intermittent medium-lengthed forwarding	0.492
Cluster 3 (8.9%)	MW-LF-MW-LF LF-MW-LF-MW LF-MW-LF-SW	medium-lengthed watching with intermittent long-lengthed forwarding	0.562
Cluster 4 (8.0%)	MB-MW-MB-LW MW-MB-LW-MB MB-LW-MB-LW	medium or long-lengthed watching with intermittent medium-lengthed backwarding	not significant
Cluster 5 (7.1%)	LW-LP-LW-LP LP-LW-LP-LW LW-LP-LW-SP	long-lengthed watching with intermittent long pause	1.257
Cluster 6 (5.3%)	MF-MF-MF-MF LF-MF-MF-MF MF-MF-MF-LF	constant medium or long-lengthed forwarding without watching	0.634
Cluster 7 (2.3%)	SP-SW-SP-SW SP-SW-MP-MW SW-MP-MW-SP	short-lengthed watching with intermittent short pause	0.477
Cluster Etc. (43.3%)	-	-	2.014

Nonetheless, the age group was a significant predictor of how much the learner rewatched a video; older adults rewatched significantly more than non-older adults do. The GEE result indicates that older adults are expected to rewatch 6.24 seconds (p ¡ 0.01) more of a video than non-older adults for a video of the same domain, level, and video length.

3.3.5 RQ2-4: How much of the video do they watch?

Interview

We identified three emerging themes that affect older adults to drop out of a video or not: (1) **circumstance** not being suitable to keep watching (e.g., time to cook), (2) **content or level** not suitable or as expected (e.g., level of yoga being too difficult), and (3) their **tendency of watching videos until the end**. Since the first two themes are obvious reasons behind dropout even among non-older adults [128], below we focus on the third theme.

Seven participants reported that they have a tendency to watch until the end and would rarely drop out in the middle: *"I always watch from the beginning to the end ... It's because, after I watch it all, I can then conclude (whether the video is useful)"* (P10), *"I always watch everything due to my desire to learn ... (Although the lecture gets boring, I watch it all) because I'm not watching the lectures for eight hours a day. I only watch for one or two hours"* (P10).

Log Analysis

We found that the age group is a significant predictor of video coverage; older adults cover significantly more of a single video than non-older adults do. The GEE result indicated that older adults are expected to cover 12.24% ($p < 0.01$) more than non-older adults for a video with the same domain, level, and video length. The distribution of the coverage also shows a similar result (Figure 3.3): for more than half of the cases, older adults covered more than 90% of the video once they started watching the video, which is much more than non-older adults. Moreover, dropping out without even watching 10% of the video, is much more common among non-older adults than older adults, taking up to around one-fourth of the non-older adults' logs.

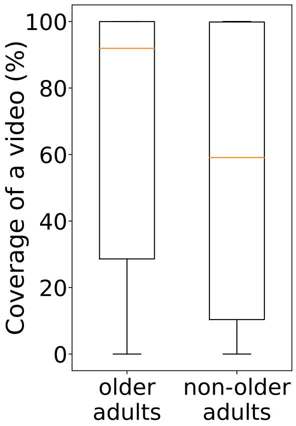


Figure 3.3: Box plot of the distribution of video coverage for both older adults and non-older adults. The orange line indicates the average coverage for each group. From the distribution, we can notice that older adults are likely to cover a video more than non-older adults.

3.3.6 RQ3: What are the difficulties they face while learning through online videos and how do they try to address the difficulties?

Interview

1. Difficulties older adults face. From the interview, participants reported that they face difficulties due to (1) the characteristics of the video medium and technology itself and (2) video-specific issues (e.g., small fonts).

Interviewees mentioned the challenges arising from the **characteristics of the video medium and technology itself**:

- **Unfamiliarity with technologies** made learning online cognitively difficult for them. Although many mentioned that they were more familiar with using smartphones and personal computers than their peers, their learning was often accompanied by inconvenience, got halted, or even restricted as they had the fear of using technology or were unable to do what they wanted to do with technology: *"I started using YouTube for less than a year ago. Previously, I thought I cannot do such things (e.g., using YouTube) at my age (so didn't even think of starting to use it)."* (P13); *"We are the generation where we used to look into printed manuals to get familiar with something. But now there's no (physical) manual and it's all stored in the phone. For us, I hope at least there's a two or three-page-long table of contents where they point out where to look at to know how to do something."* (P5).

- Participants had difficulty due to the **characteristics of the video medium itself**. Still, some of them mentioned that video is an appropriate medium considering their relatively low visual acuity and eye fatigue as (1) it is multimedia with visuals and audio and (2) missing one scene does not critically affect the overall understanding due to the context. In contrast, they also mentioned that long watching sessions are physically hard for them due to their visual/auditory ability and physical strength: *"Learning (through video) is hard since the view gets blurred and my eyes hurt after 20 minutes as I have to watch with my glasses on"* (P13). They also had cognitive difficulty interacting with the video. Three participants showed difficulty although they knew how to use the features: *"When I watch with the computer, I control using the mouse, but when I watch with small-screened phone, I just watch. I cannot control it well."* (P5). Six participants had difficulty as they did not properly know about the interaction function either partially or entirely. For example, P1 did not know about the concept of pausing or seeking so they reopened the page and rewatched the video from the beginning.

Participants also reported discomfort due to **video-specific issues**:

- **Lack of explanation on the background knowledge**. Six participants expressed cognitive difficulty due to the lack of background knowledge required for watching. They especially faced difficulty as they were unfamiliar with domain-specific jargon, the jargon used by non-older adults, loanwords, or words spoken in foreign languages. P8, who had a graduate degree, said: *"Some words are mixed with English ... I don't have difficulty (while watching a video) as I know English. But those who didn't have higher education feel the difficulty even in the streets and everywhere. They work hard to study which word is used when."*. Similarly, previous work also indicated jargon as one of the major hurdles for older adults to utilize information on the Internet [117].

- **Visual or auditory problems in videos**. They also had cognitive and physical difficulty due to visual or auditory-related problems of a video. Problems not only arise from the bad filming or editing of the video but also due to the small fonts or figures: *"They put small letters on the screen so I needed to put on my glasses. When young people do it (i.e., make a video), it's inconvenient."* (P10). Eight participants also preferred watching on a bigger screen: *"We don't have a computer. I can't buy a laptop though watching with a computer would be comfortable ... (When watching videos related to stocks through phone,) I must enlarge the chart to see the bar graphs."* (P10). Moreover, they had difficulties due to letters shown for a short time and fast speaking pace: *"Young people are fast, but as we get old we talk slow and see slow. To me, they (i.e., letters) pass away so quick that I cannot see ... They talk too fast"* (P7). Some also pointed out unclear voices: *"Some YouTubers speak in an unclear tone and some just use computer voice, but for news, we use people who can speak in a way anyone can listen to without repulsion."* (P7).

- **Distracting structure or flow**. They faced cognitive discomfort due to the structure or flow of some videos, e.g., being plain without any emphasizing. Seven participants were especially dissatisfied due to the verbose structure, irrelevant chats, and advertisements in the video. This was critical to them so they included compactness as a criterion while selecting videos to watch (RQ2-1).

2. How older adults address the difficulties. Participants (1) sought help from others, (2) searched external resources, or (3) gave up trying to resolve the issue. Participants mentioned that they **sought help from others** if, for example, they did not understand the content or they faced technical problems). Most of them sought help from their family members or acquaintances rather than interacting online through the commenting system or Q&A boards: *"(As videos are uni-directed,) ... I can only ask the tax accountant that I know."* (P12). When asked whether they read or write comments for questions, P12 said: *"Comments are just for fun. I don't think they'll be helpful to me so I don't look at the comments and just ask the tax accountant."* Several participants (5 participants) mentioned that they also tried reaching out to others beyond just their acquaintances (e.g., service center, teaching assistants) through phone calls but expressed dissatisfaction. P13 said: *"Whenever I call (the service center), the line is busy, ... and the TAs are not available. ... Once I get through the line, the TAs are very blunt while explaining ... They cannot possibly think that I don't even know this."*

In some cases, they **searched external resources** such as dictionaries, other videos/books, and the Internet when they did not know a term or content, but they rarely did the same when they faced technical problems. Only a few *searched* on the Internet, as many were not accustomed to the search function, although they mentioned that they were more familiar with technologies than others of their age.

However, there were many cases (6 participants) when they would **give up** trying to resolve the problem or just move on without trying. Some gave up because they felt it to be too much of a burden to look for the information or because they did not think that it was essential to know. Others gave up after failing to seek external help or search for external sources. P13 mentioned: *"I was going to ask my kid (on how to go back to the previous lecture), but since she seemed busy, I just moved on. ... When I ask my acquaintances, they don't really care. I know that I wouldn't understand it even though they explain it as I'm not familiar with technology. I feel bad that I didn't get appropriate help (so I had to drop the open university degree)."*

3.4 Discussions

We discuss our interpretation of the results, design guidelines for online videos and platforms targeted to support older adults' learning, and the limitations of our study.

3.4.1 Interpretation of the Results

We discuss the similarities and differences between our quantitative and qualitative results and possible explanations for the results.

Older adults want to learn subjects related to their interest or life (RQ1)

According to our qualitative results, what older adults want to learn depended more on their personal interests or needs in their life and less on jobs. This shows a clear contrast with non-older adults, as non-older adults tend to have strong career-related or educational motivations for learning [177], while older adults who are retired or about to retire may lack such motivations. This may be the reason behind our quantitative result — they watch more (1) videos on humanities and health and (2) easier videos than non-older adults (Figure 3.2). Both our quantitative and qualitative results showed that few older adults preferred to learn STEM subjects. In fact, STEM subjects were the subjects that older adults took the least compared to non-older adults (Figure 3.2). In the interview, many participants attributed the reason to the difficulty of learning STEM subjects. Considering that older adults prefer videos that (1) suit their level (RQ2-1) and (2) relate to their life, the lack of STEM videos with an easy explanation that relates to their interest could be the reason why they watch fewer STEM videos.

Older adults select videos based on the level and format (RQ2-1)

Older adults regarded the level of videos to be a critical factor while selecting videos. Our RQ1 results also support this, as they actually watch more easy-level videos and fewer hard-level videos than non-older adults (Figure 3.2). This may be due to their previous watching experiences of finding many videos to be not fitting them, which could be seen from our results: due to the level not fitting them, they dropped out of the video (RQ2-4) or faced difficulties (RQ3). Interestingly, they used the instructor's age being similar as a way to estimate the level and used it as a criterion for choosing videos. This could be partially due to the fact that an instructor in the same age group might have explained the material more easily by being in their shoes. However, it could be also because older adults may have viewed the perceived difficulty to be easier; thinking that they can also learn it just like the instructor of the same age group. Previous research also suggests that learners prefer instructors who share the same characteristics as them (e.g., gender, race, ethnicity, age) [111, 115, 165, 126].

Older adults also preferred videos that are concise and free of jokes, impromptu segments, and irrelevant chatter. This also aligns with the previous study, which argues that the preferred types of online courses are different between older adults and non-older adults: older adults prefer videos of professor lecturing, while younger learners prefer videos involving interactive learning [168].

Older adults interact less with videos and skim or skip less (RQ2-2)

Results show that older adults generally interact less with videos. This could be due to generational differences. While older adults are familiar with passive watching (e.g., TV), non-older adults including the net generation or digital natives are known to be more familiar with interactive media [144, 168].

Thus, older adults' mental model of 'online video' could be different [169]. Moreover, sequence clustering results show that they are more likely to watch in patterns that are not common (Cluster Etc.). This may be partially due to the fact that older adults are not familiar with video interactions.

Older adults tend to watch a video in a linear fashion rather than skipping forward, which also attributes to higher video coverage (RQ2-4). Research suggests that the digital environment brought changes to people's reading behaviors; non-older adults who are familiar with the web environment, which provides a vast amount of information, are likely to have a habit of skimming content [150, 178]. Our results indicate that this difference in skimming behavior is not only limited to reading but also video watching for learning.

Older adults watch videos repeatedly (RQ2-3)

Qualitative results showed that older adults watch videos repeatedly. Many participants who rewatched the video while following how-to videos wanted to first know all the steps and then follow the video. This is in contrast to how general users watch, which is by following the video in the mid of watching or segmenting the video into chunks to follow it [171]. Therefore, older adults may have watched a video repeatedly since the procedural knowledge of a whole video is beyond their working memory capacity. Moreover, older adults' pattern of watching repeatedly to fully understand the contents before starting an action may reflect that older adults tend to be reflective learners (i.e., prefer understanding things before acting) compared to non-older adults who are rather active learners (i.e., prefer getting into action and experience immediately when they are learning) [170, 144, 153, 135].

While the log analysis also showed that older adults rewatch significantly more than non-older adults, the difference was small, since the amount of rewatching was small for both age groups. This could be due to the fact that the log analysis was based on a MOOC platform, while the amount of rewatching differs according to the type of videos [102] and the format of the videos (lecture vs. tutorial) [131]. Moreover, since K-MOOC videos are organized by courses with multiple videos, this may have fostered users to proceed to the next video instead of rewatching the video. P13 reported they watch videos in a repeated manner for learning swimming on YouTube, while rarely repeating while learning English by taking courses in open university to follow the course schedule.

Older adults watch a larger portion of a single video (RQ2-4)

Our quantitative results revealed that older adults watch more parts of a video than non-older adults. Our qualitative results reveal the possible reason behind this: their tendency to watch videos until the end. This tendency to cover more parts of the video was also in line with the reason why many older adults do not want to seek forward (RQ2-2): wanting to learn without missing any parts. Another reason could be due to their watching pattern: not many watch with skimming the video through constant seek forwards compared to non-older adults (Cluster 2, 3, 6 in Table 3.5). This may explain the reason why there were many non-older adults who watched less than only 10% of the video compared to older adults. We suspect this to be also relevant with non-older adults being accustomed to bite-sized content, thus their average attention span being shorter [138].

3.4.2 Design Guidelines

Based on our results, we present design guidelines for online videos and platforms targeted to support older adults' learning.

Authoring videos that align with what older adults need

Considering topics of their interest: Our results show they like to learn those related to their personal interests, curiosity, or needs in their daily life (RQ1). Since they are relatively interested in humanities and medical subjects, **authoring diverse videos on these subjects** is needed. On the other hand, they watch fewer videos on STEM subjects (RQ1). However, this does not mean STEM videos should be created less, as there are many benefits for learning STEM subjects [130, 110]. Instead, considering their interest (RQ1) and level (RQ2-1), **more accessible STEM videos should be created that link to their interest** in health/medical domain, hobby, or life.

***Matching their level*:** Our results suggest that older adults prefer to learn by watching a video that matches their level (RQ2-1), while they tend to watch more elective-level videos than major-level videos compared to non-older adults (RQ1). Therefore, the **actual level of the material should match their desired level** of the video. This also includes giving enough explanation on the background knowledge needed, as they reported that this is one of the difficulties they face (RQ3). Moreover, efforts are needed to **decrease the perceived level of the content**. For instance, considering our results that older adults relate creators/lecturers being their age to the level of videos matching their desired level (RQ2-1), including older adults while authoring videos that can be perceived as difficult (e.g., STEM videos) can help. Another possibility is to more strictly follow Multimedia Learning Theory [154]. As the working-memory ability [123] of older adults may have decreased, following the theory to effectively utilize working-memory capacity could ultimately lower the perceived difficulty of the video.

Creating older adults-friendly videos

Making videos compact: Previous work has suggested making videos shorter is desired because users generally tend to drop out more from longer videos [142]. **Making shorter videos** is equally or even more important to older adults. This is because they are likely to face physical and cognitive difficulty from long-watching sessions (RQ3), while their video consumption pattern shows they do not seek forward or skim the video unlike non-older adults (RQ2-2). One method of making the videos compact could be to reduce jokes, irrelevant chats, or advertisements in the middle, as many value **compact information delivery** over making the video funny or interesting (RQ2-1). Especially for procedural videos (e.g., exercise video), since many older adults have a tendency to watch the video several times before following it (RQ2-3) unlike general users [171], **segmenting them into multiple videos** may help. Previous work also suggested this guideline for novice learners to allow them to freely pace the video with a pause or play interaction [104]. We suggest the same guideline applies to older adults.

Increasing accessibility of visual and auditory elements: Our results indicate that older adults undergo difficulties due to visual or auditory elements in the video (RQ3). Therefore, as suggested by the design guidelines of MOOCs for older adults [106, 166, 94, 159], **automatically adjusting the size of the visuals** can be helpful. In addition to automated adjustments, similar to mobile-friendly MOOC design guidelines [143], **enabling users to easily customize** the enlargement of visual elements in the video could be helpful. Also, from our results (RQ3), we suggest automatically **slowing down the pace of visuals** shown only for a short time. For audio, as suggested by previous work, our results (RQ3) also show that the **speed of the speech** should be slowed down [127] and **the audio quality should be high** to be heard clearly [166, 94]. Furthermore, since older adults may not be familiar with voices younger generations are relatively used to (e.g., computer-generated voice) (RQ3), enabling an **option to change the voice to the one they feel is clear** could be needed.

Increasing the delivery of the video content: Although design guidelines on the accessibility of video elements for older adults have been much investigated [106, 166, 94, 159], design guidelines on how the video content should be delivered so that videos could be more accessible to older adults have not been much explored. However, **delivering the video content better** to older adults could further increase the accessibility of the video. Our results show that older adults reported difficulty due to lack of background knowledge and language (e.g., jargon, foreign language phrases), while most do not search the Internet due to the unfamiliarity with a search engine (RQ3). Therefore, if the text or audio of a video includes such phrases, we suggest **automatically providing substitute words or relevant information links** for older adults.

Creating older adults-friendly video platforms

Recommendation engine specialized for older adults: While video recommendation takes a huge portion in how users access videos [114], the cost of recommending a video that the user would not like is higher for older adults than for non-older adults. This is because they have a higher chance to waste their time watching the video although they do not like it due to their watching pattern of being less likely to seek forward or skim the video (RQ2-2) nor drop out in the middle (RQ2-4). Especially since older adults more easily experience physical difficulties from long watching (RQ3), this cost is concerning. Thus, among various metrics to evaluate recommendation algorithms, **reducing false positives** would be a relatively more important metric when designing recommendation engines targeted at supporting older adults. To address the **cold start of recommendation** engine specialized for older adults, considering our results on what they watch and how they choose videos (RQ1, RQ2-1) can be beneficial. Moreover, similar to how Liu et al. proposed a video search interface with video accessibility metrics designed for visually impaired people [149], developing a **metric to evaluate video accessibility specialized for older adults** is necessary. We believe our results on the difficulties older adults face from video-specific issues (RQ3) could guide the design of the metric. This metric could then be augmented for recommendation engines for older adults. Furthermore, when presenting the recommendation results, different metadata could be needed for older adults: presenting **metadata that represents the video contents better**, instead of those that could only grab their attention. This is because our results show that they heavily rely on the metadata of the video for choosing videos to watch, although it often leads to misselection (RQ2-1). Therefore, metadata shown to older adults should be more carefully designed.

Providing appropriate support: Since older adults are unfamiliar with video medium, some not knowing the existence of video interaction such as seek forward (RQ3), providing **clearer instructions on the video interface** can help. Especially, since methods for video interaction are continuously evolving [163], it is necessary to understand common interaction patterns of different age groups and provide appropriate instructions that introduce the new video platform features relative to their norm of using video interfaces. Moreover, our results indicate that many older adults give up resolving the difficulty they face, which can sometimes lead to cease of pursuing learning (RQ3). Thus, channels are needed for opportune technological support or help with the content. Since most older adults did not utilize online channels (e.g., Q&A boards, chatbots) for resolving difficulties (RQ3), **offering instructions for utilizing online channels and offering offline channels** (e.g., phone calls, in-person support) is needed.

3.4.3 Limitations

There are several limitations of our study. First, we conducted interviews with Korean older adults and analyzed data extracted from a Korean MOOC platform. As cultural differences in learning can be either exaggerated [132, 156] or minimized [162, 156] as one ages, further analysis in different settings might be required in terms of generalizability.

Second, we used interaction data from the K-MOOC platform, which could be different from learning from other video platforms (e.g., YouTube). As explained in Section 3.4.1), this could have led to the difference in our quantitative and qualitative results on how older adults watch videos repeatedly. Moreover, the result of which subjects older adults take videos may be different from other video platforms. Although the K-MOOC platform provides many courses on practical content besides courses on theoretical content, the distribution of domains can be different from other platforms. Thus, the general tendency, such as older adults watching STEM domain videos less, could be not much different from other video platforms, but specific numbers of the distribution may be different.

Third, there exists a time gap between the two data streams we used for our mixed-methods approach; the log data we analyzed was collected in 2018 (pre-COVID-19), while we conducted interviews in 2020 (during COVID-19). Although previous research [183] shows that COVID-19 does not have a major impact on how learners learn a course online, COVID-19 may have changed how they watch a single video.

Lastly, our paper did not consider differences among older adults. Since there may exist differences depending on various factors (e.g., age, educational degree, gender) [180], we call for future research to investigate these factors.

3.5 Conclusion

We investigated how older adults use online videos for learning with a mixed-methods approach. We also presented design guidelines for online videos that aim to support older adults' learning. Since online videos are a prevalent medium for online learning, providing adequate support based on how older adults learn is needed to increase the accessibility of learning through online videos. We believe that our work could enable going beyond the current one-size-fits-all of online videos to better support older adults' learning.

Chapter 4. Understanding How Non-Native Speakers and Native Speakers Use LLMs Differently in Writing

Large language models (LLMs) have received significant attention due to their ability to understand and process complex natural language. This capability enabled users to easily interact by expressing their intent in everyday language, increasing the accessibility of LLMs in diverse tasks. Among diverse tasks users utilize LLMs, one of the popular uses is assisting users with writing tasks [2].

However, non-native speakers (NNS) often encounter challenges when using LLMs for writing assistance. For instance, previous work suggests that non-native speakers may struggle with formulating effective prompts due to limited vocabulary, grammatical uncertainty, or difficulties in articulating their intent clearly, leading to outputs that deviate from their desired outcomes [35, 6]. In addition, LLMs are currently performing better with English, making it harder for NNS to even use their mother tongue language to interact with LLM [20, 26, 19]. Moreover, LLMs often misinterpret prompts written by NNSs, resulting in undesirable behaviors, such as generating less accurate or even misinformative responses [34].

However, would the challenges only be limited to expressing intent? With the introduce of LLMs, writing inauthentically has been an emergent problem [1], as users just utilize the outputs of LLMs rather than appropriating them as needed with their own writing. Under the importance of writing authenticity, we try to understand how non-native speakers utilize LLMs differently from native speakers for writing and how these behaviors affect the writing authenticity.

Especially, since writing behaviors may change according to motivations or settings [25], we investigated collaborative writing behaviors with LLM by imposing the same setting on the study participants. Thus, we conducted a study in a controlled environment with 191 participants (non-native speakers (NNS) 96, native speakers (NS) 95) from Prolific, excluding insincere participation. Since low-stake nature of the task would let participants to just over-rely on LLM and finish the task as fast as possible [25], we also tried to avoid this circumstance by encouraging them to write better by giving high reward to high-quality writings. After the writing, to understand how they revise their writings, we randomly picked a sentence and asked the participants to revise the sentence given alternatives of the sentence and LLM.

After the study, we analyzed how they interact with LLM by qualitatively analyzing the user intents of the user prompts in the chat logs and quantitatively analyzing their final writings and chat logs. We found that (1) significantly more non-native speakers were likely to request a draft to LLM than native speakers while more native speakers were likely to utilize LLM for asking ideas or information, (2) non-native speakers relied significantly more on LLM's writings by directly using them in their writing than native speakers, and (3) non-native speakers revised significantly less than NS even when asked to revise while given alternatives and LLM.

4.1 Study

Here we explain how we conducted our study.

4.1.1 Study Participants

We recruited 191 participants (95 NS, 96 NNS) through Prolific to write an argumentative essay and revise it through different stages depending on the experimental condition. Participants were paid 6 GBP (approximately 7.54 USD) for the task which took about 45 minutes. There were two conditions for the revision and we recruited 102 (49 NS, 53 NNS) workers and 109 (53 NS, 56 NNS) workers for each condition. We had to eliminate participants who are NNS by their nationality but are fluent in English as NS. Furthermore, since the essay should be written by collaborating with an LLM, we had to recruit participants who has some experience using LLMs. Hence, we asked three preliminary questions before the participants started their main study.

- Q1. How well can you write in English on average? (1 = very poor, 7 = excellent)
- Q2. How long have you lived in English-speaking countries? (never, less than a year, 1-3 years, 3-5 years, more than 5 years)
- Q3. How frequently do you use Large Language Models (LLMs)? (e.g., ChatGPT, Claude, Gemini, etc.) (daily, weekly, monthly, tried, never)

The NNS participants who either answered 7 in Q1 and "3-5 years" or "more than 5 years" in Q2 as well as the the participants who answered "Never" or "Tried" in Q3 finished their study here and paid 0.1 GBP (approximately 0.13 USD). The rest of participants moved to the main study.

4.1.2 Study Procedure

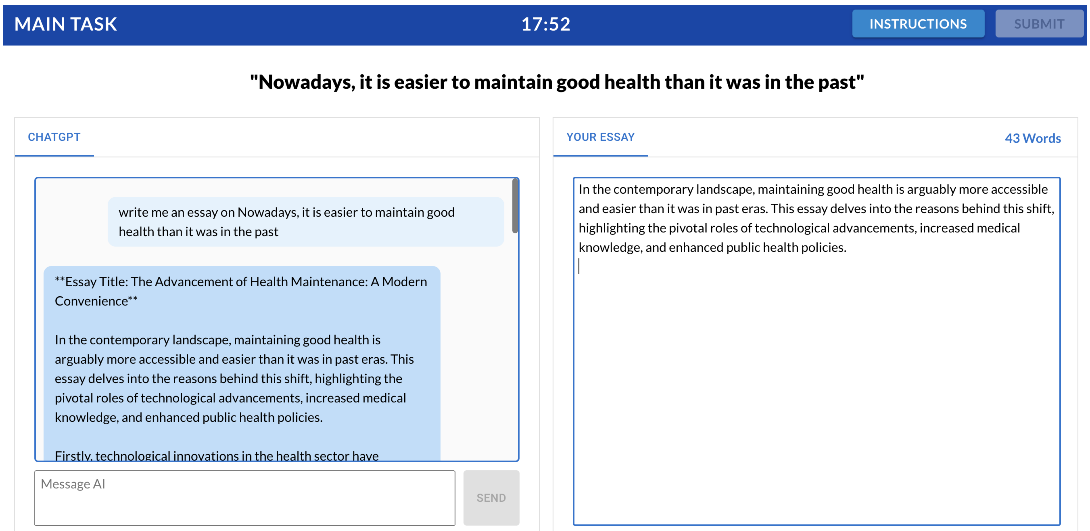
We created an interface to conduct the study on Prolific. The main purpose of this interface was to collect users' conversation logs with an LLM during writing and revising stages and collect their answers for the post-survey about their reason and satisfaction about revisions.

4.1.3 Instructions

The interface provided a set of instructions prior to writing the argumentative essay. This was given so that the participant can understand the criteria of the essay. First, the participant is given an explanation about the main task, that they have to write a 150 to 200 words persuasive article in English either for or against the given statement within 20 minutes. Second, the interface for the main writing task is explained. The participant is also informed that they can freely use the ChatGPT provided in the interface. Third, the practice writing task is explained, which would be completed before the main writing task but with shorter time limit of five minutes. Fourth, to increase the stakes of the writing task, the instruction also informed that the participants who rank in the top 40% for writing quality will receive a 100% bonus (6 GBP, approximately 7.54 USD). The grading criteria are also given here under two categories: writing style (ensure grammar correctness, use effective, diverse, and complex sentence structures, use a wide range of vocabulary and choose words/phrases appropriately, avoid incomplete writing and meet the word count criteria) and content (stay focused on the given topic, develop your main idea logically and clearly). These criteria is inspired from English comprehension tests, especially TOEFL iBT. Fifth, the participants are informed that if they finish their writing within the suggested word count, they can freely move on to the next stage. Lastly, the participant is asked to stay on the writing screen for the entire duration of the writing task.

4.1.4 Writing

As explained in the instruction, the participants are firstly asked to complete a practice writing task. This is given not only to make the participants to familiarize with the interface but also to familiarize with the writing criteria. The topic for the practice writing was “It is better to focus on tasks one excels at rather than exploring new opportunities”. The participants freely chose either for or against the given statement and wrote their essay.



"Nowadays, it is easier to maintain good health than it was in the past"

write me an essay on Nowadays, it is easier to maintain good health than it was in the past

Essay Title: The Advancement of Health Maintenance: A Modern Convenience

In the contemporary landscape, maintaining good health is arguably more accessible and easier than it was in past eras. This essay delves into the reasons behind this shift, highlighting the pivotal roles of technological advancements, increased medical knowledge, and enhanced public health policies.

Firstly, technological innovations in the health sector have

In the contemporary landscape, maintaining good health is arguably more accessible and easier than it was in past eras. This essay delves into the reasons behind this shift, highlighting the pivotal roles of technological advancements, increased medical knowledge, and enhanced public health policies.

Figure 4.1: Writing interface given to the study participants

After the practice writing task, the participants enter the main writing task. As shown in Figure 4.1, the participants can see (a) topic, (b) time left, and (c) word count of their essay in the interface. ChatGPT interface is given on the left side, and writing interface is given on the right side as shown in Figure 4.1, inspired by [25]. The topic of the main writing task was “Nowadays, it is easier to maintain good health than it was in the past”. The participants had to write 150 to 200 words persuasive essay within 20 minutes.

4.1.5 Revision

We randomly selected one sentence from their essay and asked them to revise given the embedded ChatGPT and four alternatives of the sentence generated with LLM.

We generated the alternatives in two ways. First, we modify the stylistic feature of the sentence without affecting the content that the sentence conveys. We ask LLM by employing [Prompt 1] to generate several revisions when given the sentence to revise. Second, we also modify the contents. We first ask LLM to elaborate the role of the particular sentence in full writing ([Prompt 2]), thereby obtaining the abstractive objective to focus on. Next, in [Prompt 3], we do not give the sentence but only the role, hence giving LLM some freedom to generate the sentence that is not constrained by the original content.

Prompt 1

```
Your task is to suggest 5 different writing styles of a sentence in an academic writing, without affecting its content. ’Writing styles’ mean how the ideas/content are presented.
Can you suggest 5 different writing styles of {revise_sentence+} in the below writing?
Only present 5 different sentences, not the whole writing or additional predicate.

Output should be in the format of following:
1. Sentence1
2. Sentence2

[Final Writing]
{final_writing}
```

Prompt 2

```
What is the role of ’{revise_sentence}’ in the below writing? Only present the role, not the original sentence or additional predicate.

[Final Writing
{final_writing}

Start with ’The role is to’.
```

Prompt 3

```
Based on the role of the sentence, can you suggest 5 different writings in [[TODO]] in below writing?
Only present 5 different sentences, not the whole writing or additional predicate.

Output should be in the format of following:
1. Sentence1
2. Sentence2

[Role]
{roleresponse}

[Final Writing]
{final_writing.replace(revise_sentence, ”[[TODO]]”}
```

For the revision with LLM and four alternatives, two additional questions were asked: “During revising the above sentence, which of the four suggestions did you find helpful? (You may select more than one.)”, and “How did you come up with the revised sentence? If you utilized suggestions in anyway, please describe how you utilized it. (e.g., I liked the wording “stricter regulations” in suggestion 4, so asked ChatGPT to rewrite the sentence given that phrase.)” The user can revise the sentence freely and select “None” if they did not utilize given alternatives.

4.1.6 Post Survey

After each revision of the sentences, the participants were asked to answer a post-survey question for each. If the sentence was being revised for the first time, original and the revised version of the sentence is shown and asked the following question: “For the following sentences, if you made any changes when writing Revision version, why did you not write in that way when writing Original version?”. Similarly, if the sentence was being revised for the second time (first revised with LLM and second revised with LLM and four alternatives), the first and the second revision version of the sentence is shown and asked the following question: “For the following sentences, if you made any changes when writing Revision 2 version, why did you not write in that way when writing Revision 1 version?”.

4.1.7 Satisfaction Rating

After the completing their revision and post-survey about the selected and non-selected sentences, the participants are asked to score the quality of their original and revised sentences in a 10-point scale and explain why the satisfaction changed or not. For instance, with a sentence that was revised first with LLM and second with LLM and four alternatives, the participants are asked to answer the following set of five questions:

- Q1. Please score the quality of this sentence from your original writing from 1 to 10.
- Q2. Please score the quality of first revision of the same sentence from your original writing from 1 to 10.
- Q3. Why did your satisfaction change or remain the same between the original writing and first revision?
- Q4. Please score the quality of the second revision of the same sentence from your original writing from 1 to 10.
- Q5. Why did your satisfaction change or remain the same between the first revision and second revision?

Similarly, for sentences that were revised only once, only Q1 to Q3 are asked.

4.1.8 Participant Filtering

Before analyzing the data, we first filtered out the participants who insincerely participated in the study through reading their submitted answers. We excluded those who (1) copy-pasted the same answers to all similar questions or (2) answered insincerely such as answering ’IDK’ in the open-ended questions. We ran the study remaining with 191 participants (96 NNS, 95 NS) after filtering out 9 participants.

4.2 Qualitative Analysis

We explain how we analyzed the LLM interaction patterns while writing through analyzing the collected conversation logs.

Table 4.1: Five intent categories found in LLM interaction while writing

Category	Description	Example
Draft	User asks LLM to write a full or partial draft	Write an essay about how it is easier to stay healthy nowadays than it was in the past
Idea	User asks LLM for ideas to brainstorm	give me a short list of 10 reasons why Nowadays, it is easier to maintain good health than it was in the past
	User asks LLM for information regarding the topic	Give examples for diseases that were fatal before the medical advances
Revision	After having a draft, the user asks LLM for evaluation	is the build up of the text logical?
	After having a draft, the user asks LLM for revision	Keep it in three paragraphs: intro, main argument, and conclusion. stay within word limit of 150-200 words. revise!
Synonym and Translation	User asks LLM for synonyms of certain words	synonym of maintain
	User asks LLM for translation of a text	translate opleidingen to english
Others	User prompts LLM with intents relevant to the writing task, but is not covered by aforementioned intents	how to start a conclusion but with many words

4.2.1 Main writing

We manually inspected user prompts from the main writing task. First, two authors independently reviewed the chat logs to identify recurring patterns in user intents. Through discussion, these patterns were then condensed into five distinct categories. Next, two authors independently labeled 66% of the prompts, achieving a high level of agreement (Cohen's Kappa k=0.844). Disagreements and ambiguous cases were resolved through discussion to refine category definitions. Also, a prompt may have more than two labels, or have no labels. Finally, one of the authors labeled the remaining prompts based on the finalized criteria. We present the five user intent categories in Table 4.1. Figure 4.2 shows the correlations between participants' prompts to the LLM using different intents.

4.2.2 Calculation

To analyze the distribution of the five categories, we employed two approaches: normalized counts and binary presence. For normalized counts, we examined the tendency of a user to favour certain intents over the others. For each user, we counted the prompts associated with each intent and normalized these counts to sum to one, yielding the proportion of prompts for each intent. We then computed the macro-average across the group (NS or NNS), resulting in the average proportion of prompts for each intent.

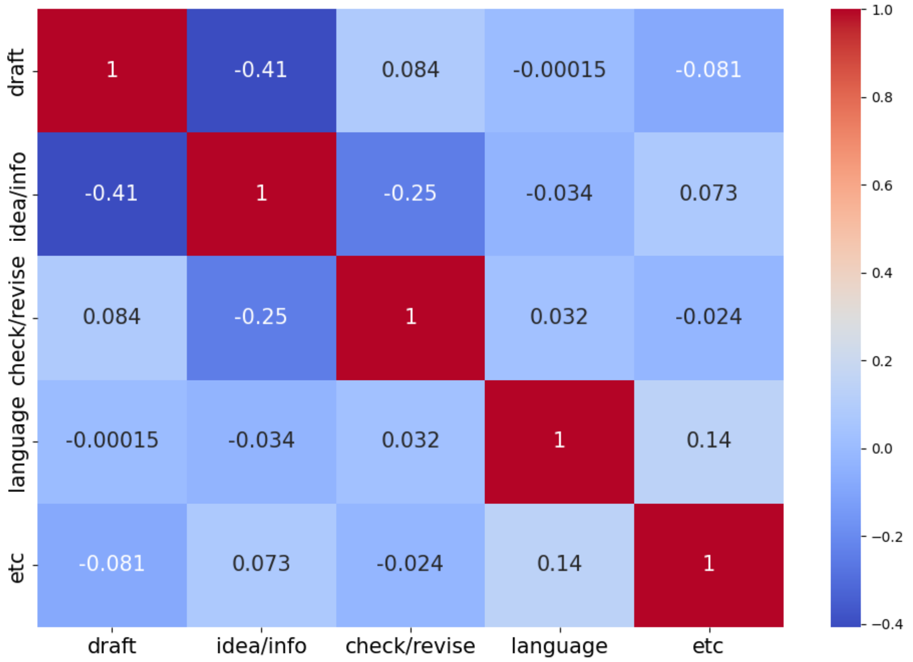


Figure 4.2: Heatmap showing the correlation between participants' prompts to the LLM using different intents

For binary presence, we assessed the likelihood of a user exhibiting an intent at least once. For each user, we checked whether they had at least one prompt associated with each intent, assigning a value of one if present and zero otherwise. We then computed the macro-average across the group (NS or NNS), resulting in the average probability of a user exhibiting each intent.

4.3 Results

We found that non-native speakers utilize LLMs differently from that of native speakers when writing. To be specific, we found that (1) significantly more non-native speakers were likely to request a draft to LLM than native speakers while more native speakers were likely to utilize LLM for asking ideas or information, (2) non-native speakers relied significantly more on LLM's writings by directly using them in their writing than native speakers, and (3) non-native speakers revised significantly less than NS even when asked to revise while given alternatives and LLM.

More Likely to Request Drafts

NNS were significantly more likely to request drafts from LLMs compared to NS. Specifically, 71% of NNS requested drafts, whereas only 50% of NS did so ($\chi^2(1, N = 178) = 7.10, p < 0.01$).

Less Likely to Ask for Idea Generation

NS utilized LLMs significantly more often for generating ideas compared to NNS. On average, NS used LLMs for idea generation 34% of the time, while NNS did so only 18% of the time (Mann–Whitney U, U = 3170.0, p <0.05).

Request more for Synonyms and Translation

NNS were significantly more likely to use LLMs for finding synonyms or translating text compared to NS. On average, NNS used LLMs for this purpose 6% of the time, whereas NS did not report using LLMs for synonyms or translation (Mann–Whitney U, U = 4473.5, p <0.01).

Relying More on LLM Responses When Writing

NNS were significantly more likely to incorporate LLM responses directly into their final writing compared to NS. We utilized maximum BLEU score between the final writing and each of the LLM responses for this analysis. We found significantly higher maximum BLEU score between the final writing and LLM responses for NNS (average: 0.48) than for NS (average: 0.30) (Mann–Whitney U, U = 5851.5, p <0.01).

Revising Less After the Writing

NNS revised their sentences significantly less than NS after the writing even when given a chance to revise while given alternatives and LLM. We utilized normalized edit distance between the sentence given to revise and the sentence they submitted after the revision as a metric for this analysis. The normalized edit distance was significantly lower for NNS (average: 46%) compared to NS (average: 64%) (Mann–Whitney U, U = 3650.5, p <0.05).

4.4 Discussion

Our findings demonstrate that non-native speakers (NNS) show different patterns when collaboratively writing with LLM — utilizing LLMs more for directly asking for the drafts to be used in their writing and not revising much after they finish writing. Previous work suggested that NNS tend to face difficulties in writing in English such as showing anxiety in English writing [9]. This could have contributed NNS to rely more on LLM-generated responses. This also aligns with the previous work that NNS are more likely to accept phrase suggestions [7] or paraphrasings [16].

Moreover, writing behaviors of NNS may have also influenced how they prompt LLMs. Previous work suggest that language proficiency influence the writing behaviors. For instance, NNS are likely to write slow [30] or pause more during their writing, especially pausing more before starting to writing the sentence [4, 36]. In addition, high-language proficiency users were more likely to plan or brainstorm their writing in the earlier phase of writing rather than in the later stage [36, 38, 11]. These writing behaviors may have led them to ask LLMs more for drafts to start their writing. NNS might approach LLMs as a means of overcoming hurdles in formulating the next sentence or starting to write without planning or structuring their thoughts in the earlier phase by requesting partial or initial draft. Moreover, LLM-generated drafts could have also lowered the anxiety during the writing as it can easily reduce the pressure of starting from a blank page.

More likeliness of asking for draft can be more severe issue as our results show they are more likely to (1) rely on the LLM's response similar to the previous work [7] and (2) fixate on the writing despite giving them AI suggestions in contrast to the previous work which suggested that they are more likely to accept AI-suggested paraphrasing [16].

The increasing popularity of LLMs can be attributed to their flexibility, allowing users to maintain control over the interaction while LLMs passively generate responses tailored to fulfill the users' intent. However, this approach may inadvertently lead to challenges for non-native speakers. This raises an important question about the responsibility for lower writing authenticity among non-native speakers. Is this solely a reflection of the choices made by non-native speakers, as they had the freedom to decide whether or not to ask for drafts and how to integrate the outputs into their writing? Or does the blame also lie with the design of LLM systems, which did not consider how non-native speakers are more likely to rely on LLMs with the current design of LLMs? This suggests a need for LLM designs that are more inclusive and mindful of the diverse ways users interact with the system, especially those with lower language proficiency which lead them to rely more heavily on LLM outputs.

4.5 Limitation

We acknowledge the following limitations of our study. First, our study setup may have influenced the results. Previous work [25] has also acknowledged that the financial compensation imposed to the writing task can influence the behaviors of participants when they write with LLM. Moreover, even with the same financial compensation and incentives, whether the participants interpret the amount can differ, which can also affect the perceived stake level of the writing task. Second, our quantitative measurement to know how much the participant relied on the LLM's response was measured by maximum BLEU score between their final writing and each of the LLM's responses, which has some pitfalls on certain cases. For instance, if the participant wrote the whole draft and then copy-pasted the draft while asking for revision and then directly submitted LLM's revised draft as their final writing, this could also result in high reliance as per our metric. Although we went over the chat logs and found that there are only few cases like this, but we still acknowledge this measurement may not correctly reflect how much the participant relied on the LLM's response in certain cases.

Chapter 5. Understanding Users' Perception Towards Automated Personality Detection with Group-specific Behavioral Data

Thanks to advanced sensing and logging technology, automatic personality assessment (APA) with users' behavioral data in the workplace is on the rise. While previous work has focused on building APA systems with high accuracy, little research has attempted to understand users' perception towards APA systems. To fill this gap, we take a mixed-methods approach: we (1) designed a survey (n=89) to understand users' social workplace behavior both online and offline and their privacy concerns; (2) built a research probe that detects personality from online and offline data streams with up to 81.3% accuracy, and deployed it for three weeks in Korea (n=32); and (3) conducted post-interviews (n=9). We identify privacy issues in sharing data and system-induced change in natural behavior as important design factors for APA systems. Our findings suggest that designers should consider the complex relationship between users' perception and system accuracy for a more user-centered APA design.

5.1 Introduction

Personality affects one's behavior in a co-located group, where all members work in the same physical location (e.g., workplaces and university labs). Personality traits, which reflect the tendency to respond in certain ways under certain circumstances [84], significantly influence job proficiency [85], job competency [52], team formation within the group [90, 78], and social dynamics in co-located group settings [75]. Thus, it has become increasingly common to conduct personality assessment of members in co-located groups and use the results to improve group productivity. One of the biggest consumers of personality tests is organizations [57] and 88% of the Fortune 500 companies have utilized a personality test [59]. As shared by an employee of a company that asks all team members to take a personality test and shares results among the team members, knowing teammates' personalities helped resolve conflicts and enhance mutual understanding (Ally Jina Kim, personal communication, April 24, 2018).

There exist diverse methods to measure personality, each with its advantages and disadvantages. Self-assessment such as Myers-Briggs [81] and International Personality Item Pool (IPIP) [66] is widely used for high applicability, low cost of implementation, and high acceptability by users [79]. However, it requires users to spare time to take the questionnaires. Automatic personality assessment (APA) tries to address these issues by predicting the user's personality by analyzing their (1) reactivity when assigned a specific simple task or (2) everyday behavioral data. However, despite the on-going debate on existence of personality change over time, APA systems that give a specific task (e.g., giving a stimulus to track eye movements [54] or introducing oneself to capture their acoustic or visual features [53]), are one-time measurements and cannot capture personality changes over time. On the other hand, APA systems utilizing everyday behavioral data, e.g., mobile phone logs [60], social media profiles [65], or wearable device logs [82], which are collected through sensing and logging technology, have the potential to measure personality continuously without direct involvement of users. However, despite the active research on APA systems with an effort to achieve state-of-the-art performance, applying these systems in practice may face resistance from users due to the use of potentially privacy-intrusive behavioral data [86]. Without a careful understanding of users' perception towards behavioral tracking for APA, such systems would pose a threat to users, hampering them from being used in the wild.

To this end, we try to understand users' perception towards personality detecting systems with everyday behavioral data in a co-located group, which we refer to as APA hereinafter. For this, we took a mixed-methods approach: a survey with 89 full-time employees and interviews with 9 participants among 32 users who experienced our research probe, which automatically detects personality with three-week-long data collection in Korea. Our findings suggest that privacy concerns in sharing data and change in users' natural behavior induced by the system during data collection are important factors to consider while building an APA system powered by behavioral data. Lastly, we provide design implications for user-centered design of automatic personality detection systems with behavioral data in a co-located group, emphasizing the importance of considering the complex relationship between user perception and system accuracy.

5.2 Methods

To understand users' perception towards APA systems that use behavioral data in a co-located group, we took a mixed-methods approach: a survey and interviews. While our literature survey suggests that multiple factors affect people's perception (e.g., privacy concerns regarding sharing the personality result, trust in result, and system-induced change in natural behavior during data collection), a single method might not provide a comprehensive view that spans multiple factors. Through the mixed-methods approach, we attempt to combine complementary insights drawn from the different methods.

With the survey, our goal is to understand respondents' acceptability of sharing their behavioral data in diverse data streams with a specific focus on privacy. Because survey respondents often have to answer questions based on their presumption rather than actual experience, we focused on questions that could be relatively easily answered based on presumption. To further gain insights into user perception based on the actual experience of using APA systems, we built a research probe informed by the survey findings. The research probe was a custom APA system that accompanies behavioral data collection of four different data streams, which varied in the level of obtrusiveness, given user control, and technology for collection (Table 5.3). After 32 participants experienced the system with three weeks of data collection, we interviewed 9 participants to investigate user perception deeper.

5.3 Methods

To understand users' perception towards APA systems that use behavioral data in a co-located group, we took a mixed-methods approach: a survey and interviews. While our literature survey suggests that multiple factors affect people's perception (e.g., privacy concerns regarding sharing the personality result, trust in result, and system-induced change in natural behavior during data collection), a single method might not provide a comprehensive view that spans multiple factors. Through the mixed-methods approach, we attempt to combine complementary insights drawn from the different methods.

With the survey, our goal is to understand respondents' acceptability of sharing their behavioral data in diverse data streams with a specific focus on privacy. Because survey respondents often have to answer questions based on their presumption rather than actual experience, we focused on questions that could be relatively easily answered based on presumption. To further gain insights into user perception based on the actual experience of using APA systems, we built a research probe informed by the survey findings.

The research probe was a custom APA system that accompanies behavioral data collection of four different data streams, which varied in the level of obtrusiveness, given user control, and technology for collection (Table 5.3). After 32 participants experienced the system with three weeks of data collection, we interviewed 9 participants to investigate user perception deeper.

5.4 Survey

We conducted an online survey (n=89) to better understand users' perception towards behavioral data sharing for personality assessment. We specifically focused on social behaviors within the context of workplace, as social behaviors commonly occur in a co-located group. We asked respondents how acceptable it is for them to share data streams within their company or organization in four aspects: (1) data collection scope across data streams (e.g., sharing online chat logs vs. offline movement logs), (2) data collection scope within a data stream (e.g., sharing online chat logs with message content vs. without message content), (3) sharing group-specific data (i.e., data that captures behaviors displayed only within a group) vs. behaviors in overall context (i.e., non-group-specific and group-specific behaviors combined), and (4) whether to have control to exclude specific data entries. Further, to understand how users' behaviors differ among the data streams, we also investigated potential differences between online and offline group behaviors, in extension to previous work [88, 73].

Differences in online and offline group social behavior patterns: We wanted to know whether there existed differences in online and offline group social behavior. We first asked how much time respondents spend on *online* and *offline* social interactions at the workplace. We chose to ask about social interaction displayed within the group as it is one of the most prevalent behaviors which can easily be found in a variety of group settings. In the survey, we explicitly gave examples of online and offline social interactions: *online* social interaction included chatting with colleagues or friends through an instant messenger and using social media, while *offline* social interaction included talking with colleagues or friends face-to-face in an informal manner, which excludes official meetings. We also asked how frequently respondents perform different social behaviors (i.e., talking a lot, starting a conversation, participating actively in a group chat, being the center of attention) *online* and *offline.*

Acceptability in sharing group-specific data with an option to exclude specific data entries: We investigated whether having an option to exclude specific data entries and sharing group-specific data (i.e., data that capture behaviors displayed only within a group) instead of sharing data in overall context can increase the acceptability in sharing data. We asked the level of acceptability in sharing certain behavioral data with their organization on a 7-point Likert scale (1-unacceptable, 7-acceptable) for each of the following three conditions of data sharing: *Data sharing condition (1)*: sharing both group-specific and non-group-specific data of a data stream, without any option to exclude data entries, *Data sharing condition (2)*: sharing group-specific data only, without any option to exclude data entries, and *Data sharing condition (3)*: sharing group-specific data only, with options to exclude specific data entries.

For the survey, we selected four online and offline data streams prevalent in modern co-located groups: online chat logs, online web or app usage logs, offline position logs, and offline movement logs.

Acceptability of sharing data across diverse data streams: We compared respondents' acceptability in sharing various data streams. We specifically focused on the data sharing level where they are asked to share group-specific data, with an option to exclude data instances (*data sharing condition (3)*). In addition to the four types of data streams (i.e. online chat logs, online web or app usage logs, offline position logs, and offline movement logs), which are easily found in modern workplaces, we also investigated their acceptability in sharing *audio* and *video* recordings, as they are richer in context yet more private. Moreover, we compared acceptability in sharing *audio* with *audio features*, such as pitch, tempo, and loudness, to reduce privacy issues within the data stream.

Acceptability of sharing data within a data stream: We wanted to understand whether sharing less information within a data stream can increase the acceptability of sharing the data stream. We asked respondents to rate the acceptability of sharing specific types of data within data streams on a 7-point Likert scale under the condition to share group-specific data, with an option to exclude unwanted data entries (*data sharing condition (3)*). Specifically, we asked in terms of three data streams as following: (1) Are people more willing to share online chat logs without message content?, (2) Are people more willing to share chat logs from public channels than private channels/DM on Slack?, (3) Are people willing to share online web/app usage data in a more abstract form (i.e., sharing URLs vs. domain information vs. categories of web/app)?, and (4) Are people more willing to share only step data than various offline movement data?

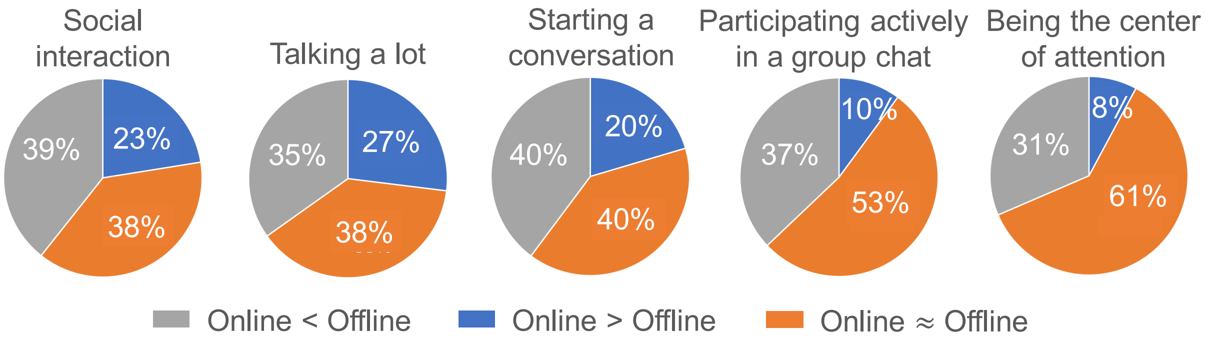
For our 50-question survey, we collected responses through various sources: an external commercial survey platform, personal contacts, and social media. As investigating differences between users with different group dynamics was beyond the scope of our research, we decided to only recruit full-time employees working in a co-located group environment to answer the survey. For quality control, we discarded respondents who spent less than 4 minutes completing our long survey (mean completion time = 22.5 minutes). From 141 initial responses, after discarding incomplete or clearly invalid answers, we ended up with a total of 89 responses (43.8% female, 33.7% aged 18∼29, 28.0% aged 30∼44, 24.7% aged 45∼60 and 13.5% aged more than 60).

5.4.1 Result

Here we report the findings from the survey.

SR1. There exists a difference between online and offline group social behavior patterns. We categorized respondents into three groups (i.e., more online-oriented, more offline-oriented, and balanced) based on the difference between frequency or time spent for each of the social interaction *online* and *offline* as in Figure 5.1. Survey results show that respondents exhibit different patterns in spending time online to offline on social interaction in the workplace: 20 respondents out of 89 (22.5%) spent more time online than offline, 35 respondents (39.3%) spent more time offline than online, and the remaining 34 (38.2%) spent a similar amount of time online and offline. Moreover, the frequencies of showing each of the social behaviors in online and offline differed.

SR2. Sharing group-specific data with an option to exclude specific data entries can increase acceptability of sharing data. Respondents overall did not find it acceptable to share any of the four data streams, with average ratings between 2.15/7 and 4.4/7 for the questions asking their acceptability to share the data stream in *data sharing condition (1), (2), or (3).* To understand how the conditions affect the acceptability of sharing data, we used Friedman's test and pairwise Wilcoxon signed-rank test with Bonferroni correction for post-hoc comparison. We observed a significant main effect of the data sharing condition on acceptability for all four data streams: online chat logs ($\chi^2(2)$=37.47), p <0.01), online web or app usage logs ($\chi^2(2)$=15.52, p <0.01), offline location logs ($\chi^2(2)$=24.17, p <0.01), and offline movement logs ($\chi^2(2)$=12.15, p <0.01). Acceptability in sharing four data streams showed a significant difference between *data sharing condition (1)* and *data sharing condition (3)* (p <0.01) as seen in Table 5.1. Respondents were negative about sharing non-group-specific data (*data*



/2/2

Figure 5.1: Online and offline social interaction patterns in the workplace: Grey represents those who show the behavior more on offline than online, blue represents those who show the behavior more on online than offline, and orange represents those who show the behavior similarly online and offline.

	Online chat logs	**Online web/app usage logs**	**Offline location logs**	**Offline movement logs**
Cond. (2)−(1)	1.36**	0.67	1.00*	0.68
Cond. (3)−(2)	0.59	0.47	0.42	0.33
Cond. (3)−(1)	1.95**	1.14**	1.42**	1.01**

Table 5.1: Difference in acceptability between data sharing conditions for each data stream (1: unacceptable, 7: acceptable). (* p <0.05, ** p <0.01)

sharing condition (1)) (online chat logs: $M = 2.15/7$, online web/app usage logs: $M = 2.35/7$, offline location logs: $M = 2.65/7$, and offline movement logs: $M = 2.52/7$), while they were more neutral about sharing only group-specific data with opt-out (*data sharing condition (3)*) (online chat logs: $M = 4.10/7$, online web/app usage logs: $M = 3.49/7$, offline location logs: $M = 4.07/7$, and offline movement logs: $M = 3.53/7$).

SR3. Acceptability of sharing data can differ significantly across the data streams. We found a significant main effect of the data stream on acceptability to share data ($\chi^2(2)$=43.89, p <0.001). Unsurprisingly, respondents would likely not accept to share audio ($M = 2.82/7$) or video recordings ($M = 2.75/7$). Sharing only audio features also showed low acceptability ($M = 3.0/7$), compared to the rest of the four data streams (online chat logs: $M = 4.10/7$, online web/app usage logs: $M = 3.49/7$, offline location logs: $M = 4.07/7$ and offline movement logs: $M = 3.54/7$). While acceptability was overall low (maximum 4.1/7), our results suggest that respondents find it significantly (all p <0.05) more acceptable to share online chat logs or offline location logs compared to audio/video recording and audio features as shown in Table 5.2. However, for online web/app usage and offline movement, there was no significant difference in sharing compared to audio/video recording as shown in Table 5.2.

SR4. With a reduced scope of the data collection, acceptability of sharing data may increase acceptability in sharing. **Online chat logs.** We asked how acceptable it would be to share all chat logs *including* message content and message metadata only (e.g., timestamp, user ID, type of message (reply or not)) *excluding* message content. We did not find a significant difference between

Web/app	*n.s.*					
Location	*n.s.*	*n.s.*				
Movement	*n.s.*	*n.s.*	*n.s.*			
Audio rec.	**	*n.s.*	**	*n.s.*		
Video rec.	**	*n.s.*	**	*n.s.*	*n.s.*	
Audio features	**	*n.s.*	*	*n.s.*	*n.s.*	*n.s.*
	Chat logs	**Web/app**	**Location**	**Movement**	**Audio rec.**	**Video rec.**

Table 5.2: Pairwise comparison of acceptability for each data stream. (* p <0.05, ** p <0.01, *n.s.*: not significant.)

the two (p = 0.52, sharing chat log including message content: $M = 3.58/7$, sharing excluding message content: $M = 3.40/7$).

Chat logs on Slack. We asked specific questions about Slack (https://slack.com), a popular workplace online instant messenger platform. Out of 89 respondents, 22 responded they have used Slack and were qualified to answer the questions. We asked about acceptability of sharing *Direct Messages (DM)* as well as messages on *Private* and *Public* channels. Public channels differ from DM or private channels since any member in the group can access the content. From the 22 Slack users who responded, we found a significant main effect of the channel type on acceptability (($\chi^2(2)$=15.70, p <0.01). Post-hoc comparisons suggest that respondents are more willing to share messages from *public* channels ($M = 4.41/7$) compared to *private* channels ($M = 2.86/7$, p <0.01) and *direct messages* ($M = 2.71/7$, p <0.01).

Online web/app usage data. We asked respondents how acceptable it would be to share (1) URLs of web pages they visit or specific app activity, (2) domain information for web pages or name of the app, and (3) only categories of web page or app (e.g., social or non-social web/app). We did not find any significant differences between these levels ($\chi^2(2)$=0.88, p = 0.65). Their acceptability in sharing online web/app usage data was all similarly low regardless of the conditions ((i): $M = 3.18/7$, (ii): $M = 3.29/7$, (ii): $M = 3.26/7$).

Offline movement data. We asked how acceptable it is to share sensor values that would indicate *movement* as well as just *steps* information. Their acceptability showed a significant difference between the two conditions ($\chi^2(2)$=5.44, p = 0.02, any movement: $M = 3.07/7$, step: $M = 3.45/7$).

5.4.2 Summary of Survey Results

Our survey results suggest that analyzing both online and offline behaviors could be effective in detecting one's personality (**SR1**). Acceptability in sharing data could vary significantly depending on whether non-group-specific data is included (**SR2**), option to exclude certain data entries (**SR2**), data collection scope across the data streams (**SR3**), and even data collection scope within a data stream (**SR4**). To further understand online and offline behavior differences and privacy concerns in depth and discover additional perceptions surrounding APA, we built a research probe by applying the findings from the survey.

Type of data	Collected content	Collected time	Data source	Intended level of obtrusiveness	Given user control	Tools used for collection
Online Messenger Usage Data	Public channel logs excluding text content	At all times	Group messenger	Low (*Only informed before the data collection*)	Can exclude data after the data collection	Slack
Online Web/App Usage Data	Timestamp of access, time spent, category of web/app	Weekdays	Inside the co-located space only	High (*Informed before the data collection, real-time/weekly report on collected data*)	Can turn off RescueTime, can exclude data after the data collection	RescueTime
Offline Location Data	Timestamp, location inside the lab in (x, y)	Weekdays	Inside the co-located space only	Medium (*Informed before the data collection, constantly giving task to make aware of data collection*)	Can turn off the watch, can exclude data after the data collection	BLE beacons, smartwatch
Offline Movement Data	Step counts, timestamp of detected step	Weekdays	Inside the co-located space only	Medium (*Informed before the data collection, constantly giving task to make aware of data collection*)	Can turn off the watch, can exclude data after the data collection	Smartwatch

Table 5.3: Summary of collected data.

5.5 Research Probe

To further understand users' perception toward APA through an actual usage experience, we built a research probe APA system. We recruited people to use the probe to understand their experience in a real context and interviewed them afterward to gain a deeper understanding of their perception. The system leverages both online and offline group-specific behaviors as people exhibit different behaviors in online to offline as found in the survey (**SR1**). It collects four different data streams (i.e., online chat logs, online web or app usage logs, offline position logs, and offline movement logs) reflecting group-specific behaviors and applying the survey results (**SR2, 3, 4**) to lower privacy concerns of participants. We also applied different levels of unobtrusiveness (e.g., giving frequent reminders throughout data collection or only in the beginning) for different data streams to better understand appropriate unobtrusiveness of an APA system. The summary of how each of the data streams is collected is shown in Table 5.3.

5.5.1 User Study with Research Probe

Our research probe involves two phases: data collection and model building. In the data collection phase, data is collected from four online and offline data streams: online messenger usage data, online web/app usage data, offline location data, and offline movement data. In the model building phase, we first extract 41 behavior features shown in Table 5.4. Then, the features are fed to a machine learning model to classify a user into one of the three classes (i.e. low, medium, and high) for each of the Big Five Personality traits [66].

We recruited four different research groups in the college of engineering at a large technical university in Korea to use our research probe. Some members of these groups chose not to participate in the study due to their own reasons. Excluding them, the four groups consisted of five, seven, nine, and eleven participants respectively, for a total of 32 participants (19% female, mean age = 26.7, S.D. = 3.7, 87.5%

Korean). Each group used a single shared space, without individual offices. The four groups varied in their culture, social dynamics, and space utilization: (1) while two groups use Korean's honorific language to communicate with each other, other groups would use it to only those who are older, (2) in one group people more closely work with each other within internal teams, while in the other groups people rather work independently on their own projects, and (3) two groups have a common area for informal social interaction while the other groups do not. Each participant received $30 for their participation in a three-week long data collection for the research probe. Institutional Review Board (IRB) approval was obtained from the university prior to the study. Participants were also asked to read and sign the terms of use, which contained information about the purpose of the study and the scope of the data collection along with study guidelines.

Phase 1: Collecting Behavioral Data

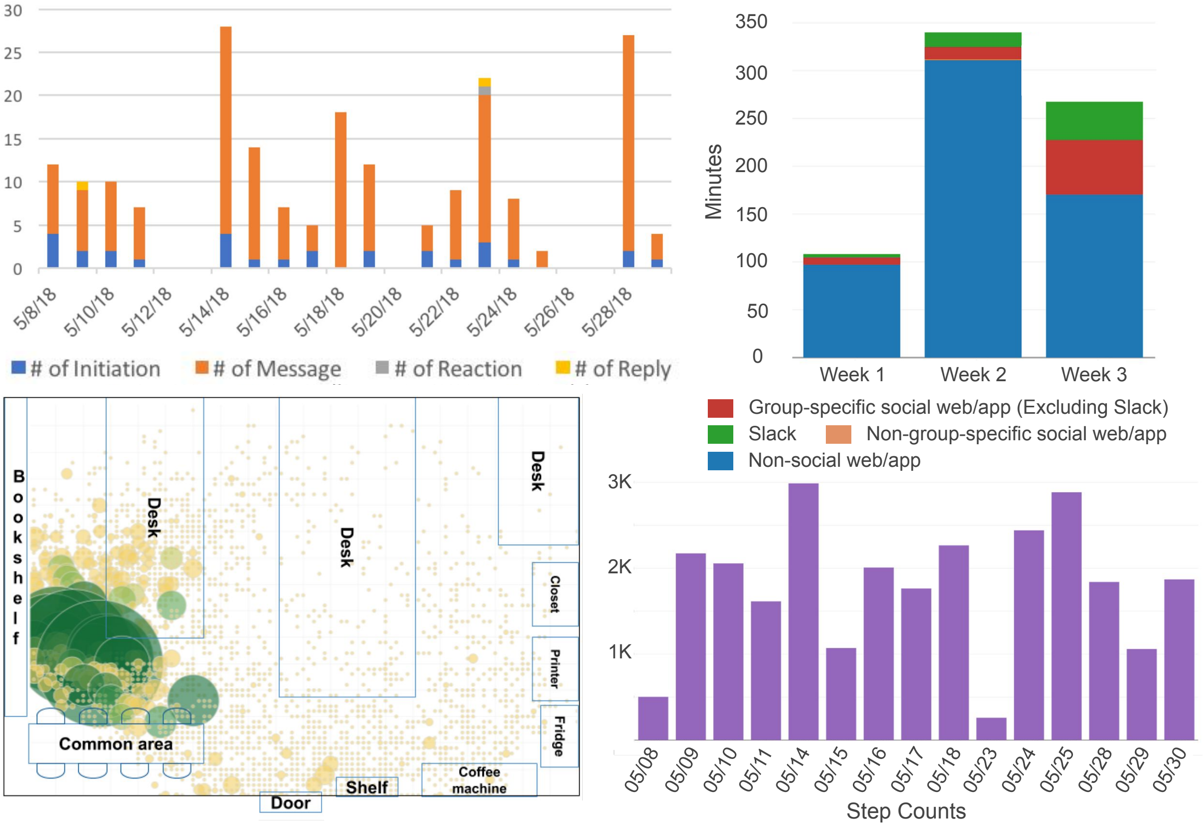
For data collection, participants were provided with a smartwatch (ASUS ZenWatch 3) and instructed to charge it whenever required and wear it. For participants who were not actively logging for several days, researchers reminded them. During three weeks in May 2018, we collected an average of 47.8 hours of offline location and movement data per person, total 2,690 online messenger activity logs (e.g., chats, reactions, participants leaving or joining a channel), and an average of 27.0 hours of online web/app usage data per person.

For the sake of transparency, after the data collection, we asked each participant to retrieve and review their online data before sharing it with us. We then provided each individual with a summary of their data as in Figure 5.2. Before analyzing the data, we gave them an option to exclude any data they want. None excluded any data instance.

To collect ground truth personality data, we asked each participant to take the International Personality Item Pool (IPIP) [66] with 100 short questions to be answered in the context of their lab to measure five dimensions of personality traits (i.e. openness, conscientiousness, extraversion, agreeableness, neuroticism) [57]. With the questionnaire result, we classified participants into three levels of each personality trait by defining the middle class as those defined with scores within one SD from the mean. This is done as the small score difference within the same class could be due to report bias. Although ground truth personality was measured six months after the data collection, it was reliable, as it showed 84.3% concurrence (27 out of 32 participants, with no participant having dramatic change of personality class—introvert changed to extrovert or vice versa) with the extraversion result which was measured right after the data collection. This is consistent with previous work that personality is relatively stable over time [58].

Online Messenger Usage Data Messenger logs contain various clues to infer one's personality [76, 70, 91], making it an appropriate source of data stream for an APA system. At the same time, it could be considered as an intrusive data stream with personal chat history. Therefore, to be less invasive, we analyze only the timestamps of messages and logs of messenger activities (e.g., message, reply, and reaction) without text content (**SR4**). We collected group messenger logs (**SR2**) of Slack, which was the communication app used in all four co-located groups that we collected data from. From the logs, we only collected public channel logs as private channels and direct messages tend to be more personal as shown in the survey (**SR2**). We also gave participants the control to discard some logs before they share the data (**SR2**). We collected messenger logs with the lowest intended level of obtrusiveness, by informing what kind of data is going to be collected only at the beginning of the study.

Online Web/app Usage Data Online web/app usage data represents digital traces of a person's



/2/2

Figure 5.2: Summary diagrams provided to the participants after the data collection. From the top left, each summary diagram represents summary of collected messenger usage data, web/app usage data, location data, and movement data.

online behaviors. However, collecting all raw traces could lead to privacy issues. Therefore, we confined the collected data to when the person was physically in the co-located group space (**SR2**) on the weekdays. Furthermore, we provided additional control options to participants to stop logging for a certain amount of time and to exclude some of the data instances (**SR2**). We collected the web/app usage data using RescueTime (https://www.rescuetime.com). Even though the survey result shows no significant difference in sharing different levels of web/app usage data, we collected domain information only for Slack. For other web or apps, we only categorized them as group-specific-social, non-group-specific-social, or non-social web/app usages (**SR4**). To discern social versus non-social web/app, we followed the classification provided by RescueTime. To discern group-specific versus non-group-specific web/app, we asked participants to fill a survey to identify whether a certain social web/app is used to interact within the group. We collected web/app usage data with a relatively high intended level of obtrusiveness: participants were continuously made aware of their online web/app usage tracking through a real-time dashboard and weekly reports provided by RescueTime.

Offline Location Data Location traces of a person can be used to infer one's personality. For example, it is shown that one's GPS logs of everyday life correlate with personality [60]. We collected participants' location information inside a physical co-located group space during weekdays (**SR2**). To collect offline location data, we developed an Android app for wearable devices that receives signals from BLE beacons (Estimote Location Beacons (https://estimote.com/products)) installed around the walls of each group's space and calculates the user's indoor position inside the space. We deployed the app in an off-the-shelf wearable smartwatch. Participants could pause data collection by turning off our data collection app or turning off the smartwatch and could exclude data instances at the end of data

Online	**Offline**
• Using (1) social-related web/app, (2) group-specific-social web/app, and (3) Slack (**E**, **N**) • Accessing (1) social-related web/app and (2) group-specific-social web/app (**O**, **A**) • Initiating a conversation on (1) any Slack channel and (2) only Slack channels including everyone • Sending / not sending a text message on (1) any Slack channel, (2) only Slack channels including everyone (**A**, **N**) • Replying to others on (1) any Slack channel, (2) only Slack channels including everyone • Reacting to others on (1) any Slack channels, (2) only Slack channels including everyone (**N**)	• Staying / not staying at one's seat (**C**, **E**, **N**) • Staying in a common area (**O**) • Going to common area (**C**) • Staying together with other group members at other than one's own seat (**C**) • Staying together with other group members in a common area • Arriving at the lab • Walking / not walking (**E**)

Table 5.4: List of group-specific behaviors that were analyzed for predicting personality. The bold alphabets represent the personality traits (O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, and N: Neuroticism) that selected the corresponding behavior as one of the top three frequently selected features throughout the cross-validation folds.

collection (**SR2**). We collected offline location data with a mid-level of unobtrusiveness: participants were constantly aware of the location data collection as they had to wear and charge the watch and were reminded to turn it on.

Offline Movement Data Movement inside a co-located group space can be indicative of one's personality trait [48, 68]. We collected movement information only within the co-located group space and excluded data over weekends to lessen the privacy concerns (**SR2**). Moreover, instead of collecting data regarding various activity information (e.g., walking, running, sitting) which can be considered intrusive by users, we only collected step count information as it is elementary information required to detect agility (**SR4**). We developed the app installed in the smartwatch to collect step counts and timestamps for each step. Participants could stop logging their movement and exclude data instances at the end of the study (**SR2**). Data was collected with a mid-intended-level of obtrusiveness with the same measure as offline location data collection: they had to wear and charge the watch and were reminded to turn it on.

Phase 2: Building an APA Model

From each of the collected data streams, we extracted 41 behavior features (19 online features and 22 offline features) as shown in Table 5.4. The full list of extracted features and how they are extracted can be found in the supplementary material. We then post-processed all behavior features to minimize the effect of differences in the group culture. For example, one group had lots of reactions added to others' messages in online messenger logs, while another group barely had any reactions. Individuals might adhere to their group culture irrespective of their own personality traits. To prevent each group's custom from influencing users' detected personality, we standardized each user behavior relative to one's own group so that every behavior feature in each group has a mean of 0 and a standard deviation of 1.

With the processed behavior features, we built an APA model to determine each participant's level of each of the Big Five personality dimensions. We ran Leave-One-Out Cross-Validation to prevent overfitting and oversampled small-numbered-classes to balance out the classes using a variant of SMOTE algorithm. Note that we performed oversampling using only the training dataset for every 10-fold cross-validation. We selected features to prevent overfitting before oversampling the training set. Then we selected the best model for each personality among a range of classification algorithms (Linear SVC, Gaussian Process Classifier, Decision Tree Classifier, Random Forest Classifier, and Gaussian NB) based on not only the high accuracy but also on F_1 macro score.

Result of the APA Model

The best model prediction accuracy for each of the Big Five Personality traits is as following: 81.3% for openness (F_1 macro score: 71.4%), 75.0% for conscientiousness (F_1 macro score: 65.8%), 81.3% for extraversion (F_1 macro score: 46.4%), 81.3% for agreeableness (F_1 macro score: 60.5%), and 71.9% for neuroticism (F_1 macro score: 58.1%). Even though the performance of our APA model suggests one possible design of automatic personality assessment with behavioral data in a co-located group, we note that we do not state the APA model itself as a research contribution in this paper, as the number of participants that we recruited is not big enough to verify its performance; we only use the result of the probe to create a realistic usage experience of an APA system to collect richer insights through the interview.

5.6 Interview

To understand users' perspective towards our research probe, we recruited 9 out of 32 participants (at least one participant from each of the four groups) who participated in the study for a post-interview. We conducted the interview in a semi-structured format for around 30 minutes and participants received $10 for compensation. During the interview, we showed them their collected data and asked their opinions regarding sharing the data for automatic personality assessment in the group. Moreover, we asked their opinion on sharing their personality result driven from their behavioral data while showing them the system prediction and their self-assessed result. We specifically focused on extraversion for this, as extraversion was a relatively widely known personality trait even among laypeople. Moreover, our focus was on sharing the detected personality from their behavioral data, not on investigating the differences in willingness to share different personality traits which were already investigated by Gou et al. [67]. Because the prediction accuracy could affect participants' perception of and trust in the system [243], we intentionally recruited both participants whose personality was predicted correctly by the system (n=5) and those with incorrect prediction (n=4).

To analyze the interview data, we transcribed the interview recordings and conducted a thematic analysis. Thematic analysis is a widely used qualitative research method to identify salient patterns or themes in the data [56]. We conducted a thematic analysis in the following five phases: (1) read transcripts while making notes, (2) go over the notes and categorize the notes, (3) tag and label themes, (4) revise tags and themes twice, and (5) re-examine tags. The first three phases were conducted by three researchers; (4) were done by two researchers; (5) was done by one researcher.

Result

The set of themes and codes resulting from our thematic analysis are presented in Table 5.5. In this section, we introduce each code in detail. In the rest of the paper, we use the shorthand codes presented in Table 5.5 to reference the codes (e.g., **BD** indicates potential benefits of using data).

Theme	Code		Example
Were participants concerned about **privacy**?	Yes	Due to **potential/imaginary** misuse (**PY1**)	*“If it is used for surveillance purpose, it will definitely be uncomfortable regardless of what. Even with the meaningless data.”* (P27)
		Due to **clear** reason (**PY2**)	*“It will feel a bit awkward to show both the part of me that I want to show and I don’t want to show (if I share my personality with others.”* (P20)
	No	Due to the **characteristic** of data/personality trait (**PN1**)	*“I feel sharing step data and messenger data are not intrusive) because . . . if I’m walking then everybody in the lab is seeing that I’m walking. And if I’m chatting in a public channel, everybody can see that I’m chatting in a public channel, everybody can see that I’m doing that”* (P25)
		Due to the **representation** of data/personality result (**PN2**)	*“(I didn’t want to erase any web/app usage data) because it doesn’t really show you much. It’s too abstract. You cannot know whether I talked to someone through Facebook messenger or whether I looked at certain page. . . ”* (P27)
		Due to other reasons (**PN3**)	*“But this kind of thing where I can collect data myself and then I can see before I share it. . . it is really important. And you can have much more trust.”* (P22)
What affects participants’ **behaviors change**?	Observer effect (**CO**)		*“(My web/app usage behaviors) could have been a bit different from when I was logging to when I was not.”* (P11)
	Self recognition (**CR**)		*“I didn’t change my behavior on Slack because I couldn’t think I was being tracked.”* (P22)
What affects participants’ **trust** in personality result?	Data-driven aspect of the result (**TD**)		*“ I was worried that I got introvert for not wearing watch”* (P28)
	Self-perception of their own personality (**TP**)		*“I thought I was mostly quiet in the lab. It was surprising that the system predicted that I was an extrovert. Basically it didn’t make sense to me”* (P30)
	Ambiguity around system (**TS**)		*“I was a bit curious whether the system predicted introverts based on online and offline data.”* (P30))
What are the potential **benefits**?	Using data (**BD**)		*“I know that that (movement) data is potentially useful for me because it seems that I’m not moving around, I should move more.”* (P22)
	Knowing personality (**BP**)		*“Data could be very useful if I want to contact somebody. So if there’s like 5 React experts in the lab, and 4 of them are introverts and 1 of them is extrovert, then I’d be more likely to ask the extrovert first than the introvert,. . . ”* (P30)
	Regarding system (**BS**)		*“If I do data collection in a long term, it is more objective (than traditional personality test). . . Also I think it is more convenient as I don’t have to think (to answer to traditional personality test).”* (P29)

Table 5.5: Results of our thematic analysis show four emerging themes: (1) privacy, (2) behavior change, (3) trust in result, and (4) benefits.

Theme 1. Were participants concerned about privacy? Overall, participants did not express much privacy concerns in the first place, even though some would state some concerns as we asked specifically. Participants’ perception on privacy varied due to various factors: scope and nature of the data, benefits provided, transparency, social desirability, and control over data collection. Some participants expressed concerns about sharing their data or personality prediction result due to: (1) potential misuse of the data other than the original purpose of data collection or intuitive concerns without a clear reason (**PY1**), and (2) clear privacy concerns towards the current scope/purpose of data collection (**PY2**). Examples of potential misuse (**PY1**) that participants mentioned were surveillance, regulating work styles, and assigning high-level meanings to data (e.g., assuming that a high step count means you worked hard). Even though they were notified of the purpose of collecting data, they would still be concerned due to possible misuse cases they could imagine on. Moreover, some expressed concerns but could not address a clear reason (**PY1**). On the other hand, participants who were concerned with a clear reason (**PY2**) pointed out different acceptability towards sharing different data streams (**SR3**): ("*(web and app were more intrusive than other data streams) because it's more personal rather than public*" (P25). Factors that affected the level of privacy concern towards a data stream were whether there exists social desirability in the data stream and whether the data stream is capturing personal behaviors rather than group-specific behaviors (**SR2**). Moreover, some expressed different levels of concerns even within the same data stream: "*not collecting the timestamp of web/app usage would alleviate the privacy issues a lot, rather collecting duration (of each web/app usage instance) would be better*" (P32). This implies that users' privacy concerns may be relieved with additional filtering within the data stream (**SR4**).

On the other hand, participants who reported no privacy concern gave the following reasons: (1) the characteristic of data or personality is limited (e.g., specific information within the data stream not being collected or being abstract (**SR4**), group-specific characteristic (**SR2**)), (2) a direct interpretation of data or personality prediction results is difficult due to the partial information that is logged instead of in a video format which shows what you did directly and the format of saved data (e.g., locations are saved in coordinates instead of a dot on a floor map), and (3) other reasons (e.g., transparent data transferring process (i.e. participants retrieving their own data to us after reviewing/deleting some unwanted instances to share for online data), trust in who they are sharing with, imperfection in data, agreement about data collection made beforehand).

Theme 2. Did the system induce participants' behavior change during data collection? Participants reported mixed responses when asked whether their behavior changed due to the system during the data collection phase. Participants who reported system-induced change of behavior said they changed their web/app usage but not others, which were deployed with high intended level of obtrusiveness as in Table 5.3. They pointed out privacy concerns arising from the observer effect (**CO**) and self-monitoring (**CR**). The observer effect refers to the unwanted change in behavior of the subject under observation due to the awareness of being observed [69]. For instance, P33 said, *"There was a feeling that someone was watching me and my behavior seems to change because of that.*" (**CR**). P29 said, *"(As I'm using web/app tracker,) I could track my web/app usage, so I wasted time (on my computer) less than usual."* (**CR**). However, many participants with behavior change from self-monitoring also reported that their behavior returned gradually. This aligns with P4's statement, who used to react to the daily summary provided as a long-time user of RescueTime, but after using it for a while, he does not anymore. Interestingly, several participants specifically pointed out that they did not change messenger usage behavior, which was collected with low obtrusiveness in contrast with web/app usage behavior. From this, we could know obtrusiveness of data collection can induce unwanted change in users' natural behaviors.

Theme 3. What affects participants' trust in personality result? Participants reported that their trust in results was affected by the following factors: (1) data-driven aspect of the system

(**TD**), (2) self-perception of their own personality (**TP**), and (3) ambiguity around how the system works (**TS**). The data-driven aspect of the system had a mixed effect on participants' trust in the personality prediction result. For instance, P25 said, *"(The result is based on) three weeks (of data). It's longer time... and it's a data-driven approach. So, I think your system is quite accurate."* In fact, for some participants, even though they were provided with predicted personality which was different from their self-assessed personality, they showed trust towards the system-driven personality. On the other hand, some participants displayed a sign of disbelief due to the limited data collection scope when presented with their result: *"I used (Facebook) through my phone and then I didn't opt to track my phone so... (I'm not sure whether enough of my behaviors are captured by the system)."* (P22). This indicates that utilization of data affects participants' acceptability towards the results. Participants' self-perceived personality also affected trust in predictions. Even though P20 and P30 both had prediction result different from their self-assessment, P20 stated, *"(The predicted) Self-assessment is as expected"*. On the contrary, P30 questioned the result and said: *"I thought I was mostly quiet in the lab. It was surprising that the system predicted that I was an extrovert. Basically, it didn't make sense to me"*. Regardless of the actual accuracy, congruence between the prediction with their self-perceived personality would affect their trust in results. In addition, some participants blamed lack of transparency on how the system works to predict personality as reasons for their distrust as in Table 5.5 (**TS**). Thus, providing the reason behind the prediction could lower their distrust.

Theme 4. What are the potential benefits? Several participants mentioned the possible benefits they could receive from (1) utilizing data itself for other purposes (e.g., reflecting/recalling on oneself's productivity, interacting with others by sharing the data, space planning based on location data) (**BD**), (2) knowing their own/others' personality (e.g., being more confident about oneself, good for new-comers to know other members, asking questions to extroverts easily) (**BP**), and (3) using APA system utilizing behavioral data over other personality measurements (**BS**). Participants reported they prefer the APA system over traditional self-assessment (**BS**) due to the convenience of easily knowing personality and 'objectiveness' in the result: *"When assessing myself, today I might feel cheerful that I might answer that I'm more sociable, but tomorrow I might be depressed. So I don't really trust it because it can result differently every day. So if data is collected for a long time and analyzed, I think that personality is more reliable. And it was more convenient that I didn't have to think a lot."* (P20). This highlights the benefit of assessing personality with users' natural behavioral data.

5.7 Discussion

In this section, we discuss design implications for user-centered design of an APA system: accuracy, privacy concerns, and system-induced change in users' natural behavior. Then we discuss the limitations of our study.

5.7.1 Implications for User-centered APA Design

Considering users' privacy concerns. In order for APA to be used in the wild, users' perception of privacy should be carefully considered. Sources of users' privacy concerns could be classified into two: (1) potential misuse or intuitive discomfort without a clear reason (**PY1**), and (2) rational thinking around current scope/purpose of data collection (**PY2**). We address possible ways to alleviate each concern. From the interview, we found that the first type of concern (**PY1**) can be alleviated by showing users the raw data to relieve users' anxiety. After showing the collected data, P29 said, "*Actually, there* *are much fewer privacy issues than I originally thought, as the content of the messages is all erased...*" This aligns with previous work on the privacy paradox [83]. Although we showed detailed terms of use—including what data is specifically collected and how—and a high-level individual summary of the collected data, they still gave users the room for imagination on what is collected. This raises *intuitive concerns* [83], although their *considered concerns* could be smaller with rational thinking on the actual scope. Thus, to minimize *intuitive concern* [83] (**PY1**), in addition to providing a high-level summary of data collection scope, providing raw data with an appropriate explanation of its use could alleviate concerns. In addition, trust in the person they are sharing the data/personality result with (**PN3**) also plays a pivotal role in users' privacy perception. P27 said, "*I don't like to share location data (with the professor) if they care whether I move around a lot.*" If users do not have enough trust in the person they are sharing with and think that they will use the data otherwise, their concerns on misuse would persist.

Another type of privacy concern arising from users' rational thinking around current scope/purpose of data collection (**PY2**), could be eased by taking preventive measures while designing an APA system. First, the characteristics of the data streams to be analyzed should be considered. Data streams that capture behaviors with clear social desirability or personal behaviors should be refrained from selection. Second, the scope of data collection even within the same data stream should be taken into account. For instance, our survey and interview results collecting only group-specific behavioral data (**SR2**) or reducing the scope of data collection within the stream to remove intrusive elements (**SR4**) could lower users' privacy concerns. Moreover, collecting the data in abstract form would be better so that raw data limits direct interpretation about the person. For example, many participants stated that their concern levels differ within the web/app data usage by the inclusion of the exact domain addresses they visited. Lastly, the measures taken to collect data also greatly influence users' privacy concerns. Giving users an option to review and exclude collected data or control to pause data collection can relieve their concern of constantly being tracked. Furthermore, the choice of technology used for the data collection could affect user's acceptability towards sharing data. For example, several participants reported that inaccuracy in offline position data alleviated their privacy concerns due to the uncertainty that is present for others to interpret one's exact position from the collected data. Participants reported that using smartwatches to share offline position information, which can be freely turned off, and beacons, which has approximately a $1m$ of error, helped reduce their perceived privacy concerns.

Moreover, users' privacy concerns regarding sharing personality results should be considered, where the factors influencing their level of privacy concerns are similar to as when they are sharing their behavioral data. Social desirability towards the same personality could be different among users: some users may believe that being more of an extrovert is less desired in workplace settings and others may think the opposite, as we found in the interview. Therefore, it is important to take into account the group culture and their interpretation of the personality traits in their unique settings.

Considering system-induced change in users' natural behavior. Unwanted change in users' natural behavior induced by the system can affect the accuracy of the system. Moreover, as users are aware of their own behavior change, their trust towards the system result can be also influenced (**TD**). System-induced change in users' behaviors is inevitable even without direct elicitation from users (e.g., giving users a task such as short presentation [53]). However, unobtrusive measures, i.e., completely not informing users, to eliminate reactivity in natural behavior can be unethical due to privacy intrusion. According to the interview, users' reported reactivity during the study was due to (1) privacy concerns arising from the observer effect (**CO**) and (2) self-monitoring caused by the high level of obtrusiveness

(**CR**). It is hard to recover users' natural behaviors if the reason behind their behavior change is privacy-related; none of the participants whose behavior changed due to the observer effect reported any sign of recovering their natural behaviors. On the other hand, if the reason is due to self-monitoring caused by the obtrusiveness of the system, users' behavior is likely to return to their natural state after a while as they get accustomed to it, which aligns with findings on reactivity [71]. Therefore, for long-term deployments of an APA system, we emphasize the importance of considering users' privacy concerns to minimize unwanted behavior change. If the data stream to be collected is privacy-wise intrusive, even keeping a high level of obtrusiveness (e.g., providing raw data or a summary of collected logs periodically) to lower privacy concerns is suggested.

Data streams to collect for better personality detection. For better personality prediction, it is important to analyze various data streams. This is because as behaviors are responses to trait-relevant situational cues [80, 55], behavior expressions may vary across various data streams due to different degrees of situational impact in these data streams. As our survey (**SR1**) suggests, one way to capture diverse behaviors is to analyze both online and offline behaviors. This can be also seen from the results of our research probe deployment: behaviors that were related to the top three selected features for the best models differed between traits as in Table 5.4. While the model for conscientiousness mostly used offline behavior features, agreeableness used mostly online behavior features, and for openness, extraversion, and neuroticism, both online and offline features were used. Hence, *analyzing both online and offline behavioral data can result in better personality detection.*

Complex relationship between accuracy and users' perception. Although it may require compromise in accuracy, it is not suggested to exhaustively collect user's data just focusing on the system accuracy. The data streams we considered in this work do not cover all trait-relevant user behaviors. Previous research suggests that minute behaviors such as voice tone [64], hand movement, and posture [50] can be useful in predicting personality. While these minute behavior could have been captured using audio/video recordings, designers of future APA systems should be careful in including these data streams, as our survey results indicated respondents' reluctance to share audio/video data (**SR3**). Respondents even reported low acceptability for sharing only certain audio features such as pitch, tempo, and loudness. Moreover, it is important to note that users' trust in results is not solely determined by the accuracy of the prediction model but also through the data-driven aspect of the system (**TD**), self-perception on their own personality (**TP**), and transparency of the prediction mechanism (**TS**). If the user changes behavior due to privacy concerns aroused from excessive data collection, their perceived accuracy could degrade. Therefore, it is rather desirable to consider the gain in system accuracy relative to the cost in users' perception for the system to be actually used in the wild. Although consideration of user perception may cause degradation in system accuracy, it should not be ignored for the successful deployment of the system.

5.7.2 Limitations

There are several limitations of our study. First, the deployment of our research probe was done in four academic research groups in Korea. As mentioned in Section 5.5.1, even though the four groups had different their culture, social dynamics, and space utilization, participants' perception might only represent users in academic environments or the Asian culture. Second, we interviewed participants by showing the predicted extraversion trained with the self-assessment result that was measured right after the data collection. Although it is different from the result shown in Section 5.5.1, it has 84.3% concurrence with no dramatic class differences. Moreover, our purpose was to understand the users' acceptability of sharing the predicted result and trust towards it for both correct and incorrect personality results as every system may have failure cases. Lastly, we used a mix of recruiting methods for the survey, which could lead to generalizability issue.

5.8 Conclusion

We investigated users' perception towards automatic personality assessment (APA) through a mixed-methods approach: a survey and interviews with participants after experiencing our research probe. We present design implications that highlight the importance of considering users' privacy concerns and system-induced change in natural behavior for designing APA systems using behavioral data in the wild. We believe that our work opens doors for more user-centered APA design to be used in the wild.

Chapter 6. Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker's Accent

Research suggests that automatic speech recognition (ASR) systems, which automatically convert speech to text, show different performances according to various input classes (e.g., accent, age), requiring attention to building fairer AI systems that would perform similarly across various input classes. However, would an AI system with the same performance regardless of input classes really be perceived as fair enough? To this end, we investigate how listeners perceive the ASR system of the same result differently according to whether the speaker is a native speaker (NS) or a non-native speaker (NNS), which may lead to unfair situations. We conducted a study (n = 420), where participants were given one of the ten speech recordings with various accents of the same script along with the same captions. We found that even with the same ASR output, listeners perceive the ASR results differently. They found captions to be more useful for NNS's speech and blamed NNS more for the errors than NS. Based on the findings, we present design implications suggesting that we should take a step further than just achieving the same performance across various input classes to build a fair ASR system.

6.1 Introduction

Automatic Speech Recognition (ASR), which automatically converts human speech to text, has allowed various technologies to foster human-to-human communication. For instance, audio conferencing tools (e.g., Zoom, Otter.ai) or video platforms (e.g., YouTube) integrate ASR technology to help users better understand each other in conversations. In the era of globalization, the ASR system is especially showing its potential in supporting communications in multilingual environments where people speaking different languages interact together with lingua franca. Despite its wide usage and practical values, ASR systems are also known for showing performance disparity regarding the speaker's various characteristics, such as accent, gender, age, and speech habit [211, 230]. For ASR technologies to be inclusive, previous research has raised the importance of reducing the performance gap.

However, would an ASR system with the same performance across various speakers' characteristics actually be fair? Previous work suggests that users' perception or reaction to the outcomes of an AI system depends not only on its performance but also on many different factors such as types of errors [220], explainability [245, 212], or user interface of the system [188]. Even if the performance is the same, an ASR system cannot be considered fair if it brings disparate impact across various subgroups [189]. Thus, to comprehend its impact, it is important to understand how listeners perceive the ASR result, considering that listeners are the primary users of the ASR systems. Although there may exist various speakers' characteristics (e.g., one's own accent, tone of the speech) which may affect how listeners perceive the ASR results, we specifically focus on native speakers (NS) and non-native speakers (NNS). This is because previous communications research suggests that speakers' characteristics, such as accents, influence how listeners perceive their speech; NNS may be perceived more negatively than NS even with the same speech content [213]. Would using ASR systems aggravate this unfavorable situation for NNS? Furthermore, would using an ASR system with the same performance for NS and NNS result differently?

To this end, we aim to understand how listeners perceive the ASR result differently when the speaker is NS and NNS with two conditions: (1) when given the same performance and (2) when there exists a performance disparity similar to the current status of ASR models. We conducted a study (n=420) where we showed a video with one of 10 speech recordings (5 NS, 5 NNS) of reading the same script along with the same caption. The participants were told that the caption was automatically generated by AI and were asked to complete a survey that asks how they (1) perceive the AI system and its output, (2) perceive the speaker and their speech, and (3) attribute the errors of the captions. We found that although the caption's performance was the same, there exist differences in how listeners perceive them. We also found that there exists an even bigger gap in listeners' perception given the performance gap similar to the current status of ASR models. Based on the findings, we present design implications for both ASR model developers and ASR system developers to build a fair and more inclusive ASR system.

The contributions of the paper are as follows:

- Findings from our study on how listeners (1) perceive AI system and its output, (2) perceive the speaker and their speech, and (3) blame the errors in the caption depending on the performance of ASR and whether the speaker is NS or NNS
- Design implications for building fairer ASR systems

6.2 Method

6.2.1 Study Preparation

Here we explain how we set up our study, in terms of how we created the 10 recordings and videos used in the study and how we created the experimental conditions.

Speaker recruitment

We recruited five native speakers and five non-native speakers of English to record the same script. Participant recruitment was conducted both offline and online. For online, we posted recruitment calls in the communities that are likely to have members from diverse countries (e.g., a Facebook group called 'foreign friends work in Korea', international communities at universities). For offline recruitment, we posted flyers on campus. We also asked applicants or international acquaintances to promote the recruitment of participants. For the application, they were asked to input their age, gender, country of origin, country they have lived in the longest, and their use of English. They were also given a short paragraph to read aloud and were asked to record it. We recruited 10 participants by mainly considering (1) their demographics to balance the age and gender between the two groups, and (2) nationality and their region so that the participants are from different parts of the world. Since one's own accent may differ significantly even within a country [235] for non-native speakers, we tried to select those with strong representative English accents of their own country. For this, we selected NNS who answered the country they lived in the longest corresponded to their nationality. Moreover, since some applicants had weak English accents, two or more authors went over the candidates and selected those whom they both agreed on having strong English accents. Table 6.1 shows the information of the speakers. We also asked if they had other factors that could potentially affect their speaking ability. Although NNS1 said ADHD, they said it does not affect their pronunciation, but just influences the duration of getting the recording task done. Moreover, NNS4 reported wearing braces, but since wearing braces can be common, we proceeded with the recording.

Audio recording

For each of the 10 selected participants, we proceeded with a remote recording session, which lasted approximately two hours. To avoid external noise during the recording as much as possible, participants were asked to use a microphone and computer in a quiet place during the session. Since the listener's perception towards the speaker and the ASR system may differ greatly according to various factors (e.g., level of the topic, word selection in the speech, grammatical errors), all speakers were given the same script. The script was based on the lectures from a Coursera course 'Understanding the Brain: The Neurobiology of Everyday Life' [1] and was modified so that it contains no grammar errors, contains no domain-specific jargon, and the flow is smooth. The script was about introducing the four functions of the brain: voluntary movement, perception, homeostasis, and higher cognitive abilities. The script consisted of 843 words (10 paragraphs). The speakers recorded under the presence of at least one author and were asked to record at least three times per paragraph using their own pronunciation and voice. If they made mistakes during the recording, such as mispronouncing a word, correcting themselves, or unnatural hesitation, they were asked to say "one, two, three" and then re-start recording the sentence from the beginning. To ensure the quality of the recordings, at the recording session, at least one author went over the recordings and asked for re-recording if there still remained any problem. After they finished the recording, we chose the version with the fewest external noise and unintended pauses between the words compared to the other versions recorded by the same speakers.

Audio post-processing

We post-processed the voice recordings, such as adjusting the average volume of the audio to ensure consistency across all speakers' recordings and experimental conditions. We also trimmed out the mistakes and instances of "one, two, three". Furthermore, we concatenated the paragraph-by-paragraph recordings into one recording, setting the time interval to about two seconds between them so that the listeners could tell that the paragraphs were broken up. We removed any non-voice noise such as background noises at the beginning or end of each paragraph recording.

Experimental Conditions

We had three caption conditions to mimic the three levels of performance of ASR: Word Error Rate (WER) 5%, WER 15%, and WER 30%. We decided on these three WERs considering the current ASR technology performance from Section 6.3.1 as well as the literature on it. Research suggests that humans can perceive differences in WER that are greater than 5-10% [236] and that 20% is the critical point for the transcription to be useful and acceptable [194, 224]. Therefore, we set our caption conditions as WER 5%, 15%, and 30%.

Even with the same WER, the perceptions of how listeners perceive the caption could be different depending on the error type in the caption [244]. Therefore, we made the same caption with each of the WER to be used across all of the recordings.

To generate captions with a target WER, we utilized the errors in generated captions from various ASR technologies 6.2. If the errors appear in the captions of multiple speakers and multiple ASR

[1] https://www.coursera.org/learn/neurobiology

Table 6.1: Demographics and information of speakers (5 NS, 5 NNS) who participated in audio recording

Participant		English as Mother Tongue	Nationality	City and Country Lived the Longest	Age	Gender	Fluent in English*	Close with Accented English Speaker*	Use Accented English*	Use English Frequently*	Speak English Frequently*	Other Factors Affecting Speaking Ability
NS	**1**	Yes	USA	California, USA	22	Female	7	4	2	7	7	None
	2	Yes	UK	London, UK	23	Female	7	7	1	6	6	None
	3	Yes	New Zealand	Auckland, New Zealand	41	Male	7	7	1	7	7	None
	4	Yes	Australia	Sydney, Australia	24	Female	7	7	3	7	7	None
	5	Yes	UK	Liverpool, UK	25	Male	7	7	7	7	7	None
NNS	**1**	No	Greece	Athens, Greece	30	Male	6	1	6	7	7	ADHD
	2	No	Ecuador	Quito, Ecuador	25	Female	6	6	6	6	6	None
	3	No	Tajikistan	Panjakent, Tajikistan	25	Female	3	6	7	3	3	None
	4	No	Vietnam	Binh Thuan Province, Vietnam	25	Female	6	6	6	6	6	Wearing Braces
	5	No	China	Anshan, China	23	Male	3	5	6	5	3	None

*The answer was given as 7-point Likert scale (1 = very strongly disagree, 7 = very strongly agree)

technologies, it is likely that the error is plausible (e.g., for the error that was originally 'while' but got translated to 'well' by ASR, was shown in six out of ten different speakers, seven out of seven different ASR models). Therefore, we selected frequent errors until two-thirds of the target WER was reached. For the remaining one-third of the target WER, we selected errors by considering the distribution of already selected errors. This is because if the errors are concentrated in a certain part of the recording, the participants may perceive the quality of the transcription differently. Thus, to make sure that the errors are evenly distributed throughout the recording, we partitioned the transcript into micro level (three words per partition; total 281 parts) and macro level (approximately 84 words per partition; total 10 parts) and set the entropy of the sum of substitutions (S), insertions (I), and deletions (D) to be higher when selecting errors to include in the caption. We also kept the capitalization, punctuation marks, and abbreviations (e.g., it's, they're) unchanged from the original script unless they are interpreted differently by ASR technologies.

In this way, the total number of words from the original captions may change. To make sure that the changed caption also appears as the speaker speaks, one author first created ground truth captions per recording by syncing the original script with the timestamps of the speaker's utterance as a subscription text (SRT) file, and then we synced the changed caption to the ground truth caption. We synced the changed caption so that the same number of words can newly appear in each caption. If the number of words does not divide evenly, we made the remaining words appear in later parts, so that the captions do not appear faster than the voice. Finally, to give the impression that the caption is auto-generated by an ASR system, we delayed the caption timestamps by 0.2 seconds and explicitly told the participants that the caption was auto-generated by an AI.

Video processing

We created videos that experiment participants would watch, with the audio of the processed audio and captions of a given condition. Since the perception towards the speaker may be largely influenced by the appearance of the speaker, we created videos of black screen. But to make sure that the listener is looking at the video and to filter out those who have not paid attention to the video, we also inserted four pictures of animals at approximately 1/5, 2/5, 3/5, and 4/5 points of the video for 10 seconds each. The images were selected so that each could leave a strong impression so that the participants could pass the attention check question without much effort.

6.2.2 Study

Here we explain how we conducted our study.

Study Participants

We recruited a total of 420 participants through Prolific [2] to watch a video with the captions of a certain experimental condition. Participants were paid 3.75 GBP (approximately 5 USD) for the task which took about 25 minutes. For each condition (total 30 conditions = 10 videos (5 NS, 5 NNS) × 3 levels of caption conditions (WER 5, 15, 30)), we recruited 14 listeners (7 NS, 7 NNS).

Since the videos were in English, we only recruited participants who could understand spoken English — who answered more than 4 for the questions on how much they can understand spoken English in daily conversation as well as in academic purposes with a 7-point Likert scale (1 = cannot understand at all, 7 = can fully understand).

Study Procedure

We created an interface to conduct the study on Prolific. The main purpose of this interface was to assign a condition for each participant, allow them to watch the video in a controlled environment, and collect their answers for the post-survey regarding their experiences and perceptions.

Video Assignment The listener's familiarity with the accent may influence how the listener perceives the ASR system and the speaker [185]. Since people are relatively more familiar with NS's accents while less familiar with NNS's accents in the real world, (Section 6.3.2), we assigned NS's video if the listeners are familiar with the NS's accents or NNS's video if the listeners are not familiar with the NNS's accent.

For this, participants were asked to provide their familiarity with the 10 accents of the speakers in the videos. The participants had to listen to each sound clip, which was around 10 seconds long, and provide their familiarity with the accent on a 7-point Likert scale (1 = not familiar at all, 7 = extremely familiar). These sound clips were chosen based on internal consensus; three of the authors picked multiple clips from each video that represented the characteristics of the accents of the speakers. Among the clips that two or more authors agreed on, the final clips were decided so that all speakers did not share a common line of the script. This was done since listening to the same script multiple times could influence how well the participant hears the sound clip, potentially affecting participants'

[2] https://www.prolific.co/

familiarity with the accent. To prevent any ordering effects, the sound clips were provided in random order.

Based on their familiarity with the speakers' accents, we randomly assigned a NS's video that they had rated 5 or higher, or a NNS's video that they had rated 3 or lower. Hence, if a participant rated their familiarity higher than 4 for all NNS and less than 4 for all NS or if all of the videos that can be assigned are fully assigned already (i.e., 14 participants per each condition who fully completed the task and is not filtered out (Section 6.2.3)), they completed the task at this point and were paid 0.9 GBP (approximately 1 USD).

Instructions The interface provided a set of instructions prior to watching the video. This was given so that the participant could prepare the appropriate environment to watch the video.

First, the participants were asked to play a 5-second audio clip with a beep sound to unmute or adjust their audio to be able to clearly hear the sound of the assigned video. Second, the main task was explained. The participants were told that there would be a video with captions auto-generated by the AI system. They were warned that the video interactions (i.e., pause, skip forward, skip backward, re-watching) would be disabled once the video started playing. This was done so that the listeners only get exposure to certain parts of the caption once. Furthermore, they were warned that they must not refresh the page as this could also allow them to re-watch the video. Third, they were asked to concentrate on the video, as they would be asked to recall some content of the video later for attention-checking purposes. Similarly, they were also told to concentrate on the screen as animal images would randomly appear in the video, which would also be asked later. This was to make sure that the participants actually listened to the video and watched the screen. However, we did not ask them to always look at the caption so that it could mimic the real-world situation, where they get to switch back and forth between the screen and the captions in their free will.

Video Watching After the instructions, the participants were shown a video of the assigned condition. Although the speakers all read the same script, the length of the video slightly varied depending on their speaking speed. The average length of the video was 363.5 seconds (SD = 77.4, min = 270.0, max = 494.0). As instructed, all video controls (pause, skip forward, skip backward, re-watching) were disabled. Any action of refreshing the page was recorded if it occurred. A separate progress bar was inserted at the bottom of the video to show the progress so that the participants can track where they are at and how much is left. This was done to help the participants maintain their concentration. After the video ended, they could move on to the post-survey.

Post-Survey We asked the following main questions to answer our RQs:

- RQ1. [Perception towards ASR system and its output]
  - How would you rate the quality of the AI technology that generated the captions you watched in the video? (1 = very poor, 7 = excellent)
  - Captions were useful for recognizing the speaker's pronunciation. (1 = strongly disagree, 7 = strongly agree)
- RQ2. [Perception towards the speaker and their speech]
  - How much did the speaker's accent negatively affect your understanding of the video content? (1 = never, 7 = every time)

  - The speaker is highly knowledgeable about the subject matter (1 = strongly disagree, 7 = strongly agree)
  - I found the speaker's explanation to be reliable and trustworthy. (1 = strongly disagree, 7 = strongly agree)
  - The speaker is highly skilled in delivering the content. (1 = strongly disagree, 7 = strongly agree)

- RQ3. [Error blaming]
  - How much of the errors in the caption do you think were caused due to the speaker? (e.g., speaker's accent, way of talking, or speaking habits) (1 = never, 7 = every time)
  - How much of the errors in the caption do you think were caused due to the AI? (e.g., AI's low performance of recognizing speech) (1 = never, 7 = every time)

Since questions on error blaming could not be answered if the participant did not notice any errors in the caption, we asked a question whether they noticed any errors, and if they answered "no", error-blaming questions were skipped. Furthermore, through pilot studies, we found that some could not answer questions on the expertise of the speaker and the reliability of the speaker's explanation, we provided an option to skip the question.

For quality control, participants were asked to select an animal image that did not appear in the video. Moreover, they were asked to briefly write the topic of the video.

6.2.3 Participant Filtering

Before analyzing the data, we first filtered out the participants who failed to pass our quality control questions or who were not suitable for our analysis. The criteria to exclude were:

- The participant refreshed the page once or more.
- The participant was incorrect on the quality-control question regarding animal images from the video.
- The participant was completely wrong in describing the topic of the video.
- The participant did not pay attention to the caption (answered below 3 in question on 'How much attention did you pay to the captions while watching the video? (1 = never, 7 = every time)').
- The participant did not notice any errors in the caption.
- The participant did not show an opposite trend for the same question but when asked in an opposite manner.

We ran the study until each of the 30 conditions gathered responses from 14 participants (7 NS, 7 NNS) after filtering out using the above criteria, resulting in a total of 420 participants. Out of 580 participants who completed the task, 160 were excluded.

6.3 Result

In this section, we first present our analysis of the performance of eight different ASR models. Then, we present the distribution of familiarity towards NS and NNS's accents. Finally, we present the results to understand how listeners perceive the ASR system (RQ1) and the speaker (RQ2) and how they attribute the errors (RQ3) differently when the speaker is NS and when the speaker is NNS. We present these results under two circumstances: (1) given the same ASR performance regardless of the speaker being NS/NNS and (2) given the disparity gap which reflects the current status of ASR systems as in Section 6.3.1.

6.3.1 Performance of ASR Model

We used the 10 recordings (5 NS, 5 NNS) that we collected and processed (Section 6.2.1 and 6.2.1) to understand and compare the performance of various ASR models available: (1) three commercialized ASR technology integrated into video platforms or meeting support systems, namely YouTube automatic captioning function [3], Zoom automated captions function [4], and Otter.ai [5]), (2) four ASR APIs, namely Rev AI [6], AssemblyAI [7], Amazon Transcribe [8], and IBM Watson [9]), and (3) one of the state-of-the-art ASR models, namely OpenAI's Whisper [228] in two versions (English-only model and multilingual model, both in base size).

Results (Table 6.2) show that the performance of ASR models measured by WER varies significantly across different speakers and models: 0.5% (AssemblyAI, Speaker 1 (NS)) to 100% (YouTube, Speaker 10 (NNS), failed to output ASR result). The performance of a single ASR model also varied a lot (std WER of YouTube: 30.3%).

Despite the performance differences across speakers, we could still find a clear tendency for the ASR performance of NNS to be lower than that of NS overall. The average WER of NS varied from 1.3% to 29.8%, while the average WER of NNS varied from 8.0% to 65.6%. The ASR performance was also more stable across NS than NNS (std for NS: 1.2% to 13.8%, NNS: 4.5% to 40.4%). For one NNS, YouTube even failed to output the ASR result, resulting in 100% WER. This result of the disparity gap between NS and NNS also aligned with previous work [191].

Interestingly, all three commercialized online platforms that integrate ASR technology (YouTube, Zoom, Otter.ai) and IBM Watson showed huge performance disparity (approx. 15%) between NS and NNS compared to other models. In addition, considering 20% WER to be the critical point for ASR models to be useful [194, 224], only two of the commercialized ASR technology (YouTube, Zoom) showed the performance of being useful for NS, while none of them were useful for NNS. This could mean that users in the wild would be perceiving a larger performance disparity between NS and NNS compared to what is being reported in previous research or API documents.

[3] https://support.google.com/youtube/answer/6373554?hl=en

[4] https://support.zoom.us/hc/en-us/articles/8158289360141-Enabling-automated-captions

[5] https://otter.ai

[6] https://www.rev.ai/

[7] https://www.assemblyai.com/

[8] https://aws.amazon.com/transcribe/

[9] https://www.ibm.com/cloud/watson-speech-to-text

Table 6.2: Average and Standard Deviation of Accuracy (WER) of Five NS and Five NNS for each ASR model

		YouTube	ZOOM STT	Otter.ai	Rev AI	AssemblyAI	Amazon	IBM Watson	Whisper Base.en	Whisper Base
NS	AVR	2.5	8.0	4.3	2.9	1.3	2.1	29.8	2.5	3.4
	STD	2.4	7.1	3.9	2.9	1.2	1.9	13.8	1.6	3.0
NNS	AVR	28.64	24.6	19.1	14.8	8.0	9.0	65.5	11.2	17.0
	STD	40.4	12.1	12.1	7.2	5.4	4.5	19.1	6.2	13.9

6.3.2 Familiarity towards NS and NNS's accents

Since the listener's familiarity with the speaker's accent may affect the listener's perceptions, we analyzed people's accent familiarity to take this factor into account when assigning conditions in the study. We analyzed 875 participants' (405 NS, 470 NNS) responses to accent familiarity for each of the 10 recordings (5 NS, 5 NNS) on a scale of 1 (not familiar at all) to 7 (extremely familiar). These participants included the main study participants (420 total), as well as participants who (1) dropped out after completing accent familiarity ratings, (2) could not proceed to the main task as the videos that could be assigned were already fully assigned to 14 participants, and (3) got excluded later in the main study (Section 6.2.3). There were no participants who could not proceed to the main task due to rating their familiarity higher than 4 for all NNS and lower than 4 for all NS.

Figure 6.1 shows the distribution of accent familiarity towards speakers differed depending on whether the speaker was NS or NNS. Participants were relatively **more familiar with NS's accents** and **less familiar with NNS's accents**: on average, 48.9% of participants were familiar (responded with 4-7) and 5.09% of participants were unfamiliar (responded with 1-3) with NS's accents. While 34.22% of participants were familiar, 88.3% of participants were unfamiliar (responded with 1-3) with NNS's accents.

We also conducted Pearson's Chi-squared test and found that there is a significant relationship between accent familiarity and whether the speaker is NS or NNS (χ^2 = 3491.5, *df* = 6, p ¡ 0.001). Cramér's V, which shows how strongly two categorical variables are associated, was 0.63, meaning that the two variables have a strong relationship. Including variables with a strong relationship in the statistical model may result in multicollinearity, which results in high errors in the analysis result. Thus, we decided to exclude accent familiarity as a factor in the following statistical analysis as suggested by Dormann et al. [197].

At the same time, accent familiarity differed even within the NS and NNS groups. For instance, participants were relatively less familiar with NS5's accent compared to NS1's accent, shown in Figure 6.1.

6.3.3 Perception difference when the performance is the same

RQ1: Perception towards ASR system and its output

Quality of ASR system We performed Aligned Rank Transform [239] to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners perceive the quality of the ASR system (Figure 6.2-left). We found that the performance of ASR has a statistically significant effect ($F_{2,417}$ = 86.07, p ¡ .001, η_p^2 = 0.29), while **whether a speaker is NS/NNS did not have a significant effect on how listeners perceive the quality of the ASR system** ($F_{1,418}$ = 3.49, p ¿ .05). We also did not observe a significant interaction between these two factors ($F_{2,414}$ = 1.77, p ¿ .05).

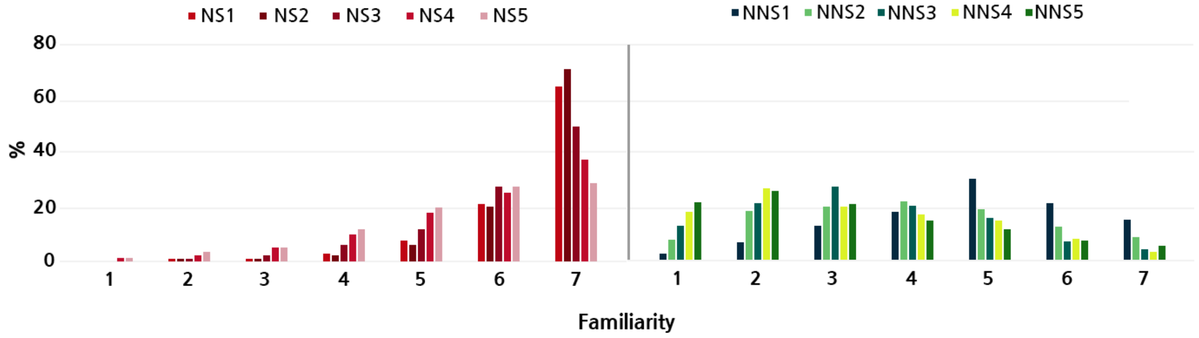


Figure 6.1: Listeners' familiarity with accents of NS and NNS

For the post-hoc pairwise comparison, we used ART-C contrasts with Tukey adjustment [200]. Results showed that listeners **perceived the quality of the ASR system to be significantly better as the WER gets lower** in our experimental conditions: WER 5 ($M = 5.26$, $SD = 1.03$) ¿ WER 15 ($M = 4.60$, $SD = 1.21$, p ¡ .001), WER 5 ¿ WER 30 ($M = 3.37$, $SD = 1.44$, p ¡ .001), WER 5 ¿ WER 15 (p ¡ .001).

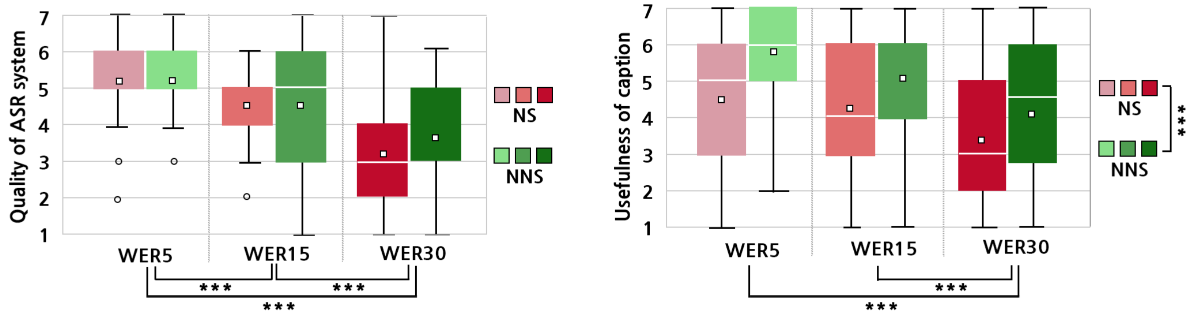


Figure 6.2: Listeners' perception towards the ASR system and its output (*** p ¡ .001)

Usefulness of captions We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners perceive the usefulness of the caption (Figure 6.2-right). We found that whether a speaker is NS/NNS ($F_{1,418} = 39.24$, p ¡ .001, $\eta_p^2 = 0.09$) and the performance of ASR ($F_{2,417} = 22.42$, p ¡ .001, $\eta_p^2 = 0.10$) have a statistically significant effect on how listeners perceive the usefulness of the caption. However, we did not observe a significant interaction between these two factors ($F_{2,414} = 0.14$, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [200]) results showed that listeners' **perceived usefulness of captions was significantly higher when the speaker is NNS** ($M = 5.05$, $SD = 1.69$) compared to when the speaker is NS ($M = 4.10$, $SD = 1.91$, p ¡ .001). Post-hoc tests also showed that listeners **perceived significantly lower usefulness for WER 30** ($M = 3.77$, $SD = 1.85$) **compared to WER 15** ($M = 4.74$, $SD = 1.77$, p ¡ .001) **or WER 5** ($M = 5.21$, $SD = 1.67$, p ¡ .001).

RQ2: Perception towards the speaker and their speech

We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on the perception towards the speaker and their speech (Figure 6.3). We found that

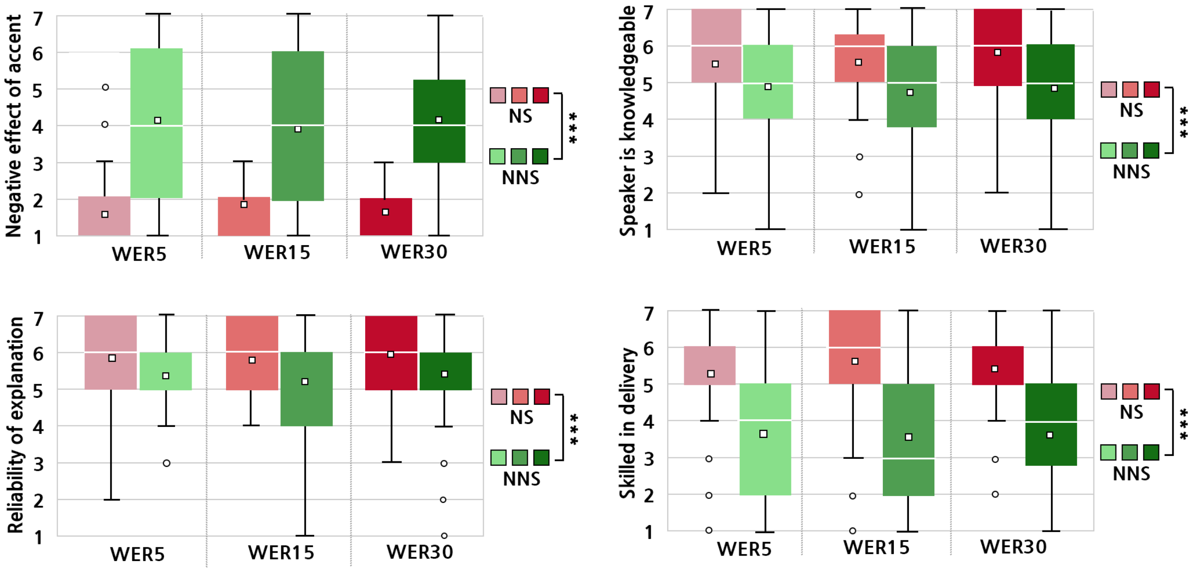


Figure 6.3: Listeners' perception towards the speaker and their speech (*** p ¡ .001)

whether a speaker is NS/NNS has a statistically significant effect on the perceptions towards the speaker and their speech: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,418}$ = 312.34, p ¡ .001, η_p^2 = 0.43), perceived expertise of the speaker ($F_{1,418}$ = 25.67, p ¡ .001, η_p^2 = 0.06), reliability of speaker's explanation ($F_{1,418}$ = 36.10, p ¡ .001, η_p^2 = 0.09), and perceived delivery skill of the speaker ($F_{1,418}$ = 195.64, p ¡ .001, η_p^2 = 0.32). On the other hand, **performance of ASR did not have a significant effect on the perceptions towards the speaker and their speech**: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,417}$ = 0.05, not significant), perceived expertise of the speaker ($F_{1,417}$ = 0.29, not significant), reliability of speaker's explanation ($F_{1,417}$ = 0.14, not significant), and perceived delivery skill of the speaker ($F_{1,417}$ = 0.10, not significant). We also did not observe a significant interaction between these two factors: perceived negative effect of the speaker's accent on their understanding of the video content ($F_{1,414}$ = 1.02, not significant), perceived expertise of the speaker ($F_{1,414}$ = 0.10, not significant), reliability of speaker's explanation ($F_{1,414}$ = 0.09, not significant), and perceived delivery skill of the speaker ($F_{1,414}$ = 0.48, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [200]) results showed that **listeners perceived the speaker and their speech significantly negatively when the speaker is NNS** compared to when the speaker is NS: perceived negative effect of the speaker's accent on their understanding of the video content (p ¡ .001, NNS: M = 4.08, SD = 1.77, NS: M = 4.08, SD = 1.77), perceived expertise of the speaker (p ¡ .001, NNS: M = 5.66, SD = 1.72, NS: M = 4.87, SD = 1.30), reliability of speaker's explanation (p ¡ .001, NNS: M = 5.35, SD = 1.28, NS: M = 5.90, SD = 0.97), and perceived delivery skill of the speaker (p ¡ .001, NNS: M = 3.66, SD = 1.65, NS: M = 5.49, SD = 1.32). This result aligns with previous works that also found that NNS receive unfavorable impressions compared to NS [201].

RQ3: Error blaming

Blaming the ASR system We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners blame the ASR system for the errors in the caption (Figure 6.4-left). We found that whether a speaker is NS/NNS ($F_{1,418}$ = 74.30, p ¡ .001, η_p^2

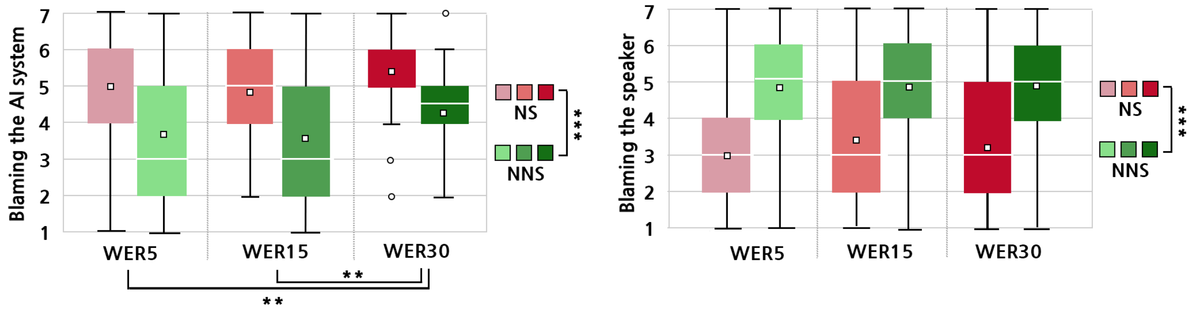


Figure 6.4: How much listeners blame errors in the captions on the AI system (left) and the speaker (right) (** p ¡ .01, *** p ¡ .001)

= 0.15) and the performance of ASR ($F_{2,417}$ = 9.75, p ¡ .001, η_p^2 = 0.05) have a statistically significant effect on how much listeners blame the ASR system for the errors in the caption. However, we did not observe a significant interaction between these two factors ($F_{2,414}$ = 0.19, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [200]) results showed that listeners **blamed the ASR system significantly more when the speaker is NS** (M = 5.10, SD = 1.53) compared to when the speaker is NNS (M = 3.86, SD = 1.69, p ¡ .001). Post-hoc test results showed that listeners **blamed the ASR system significantly more in WER 30** (M = 4.89, SD = 1.53) than WER 15 (M = 4.26, SD = 1.70, p ¡ .01) or WER 5 (M = 4.28, SD = 1.86, p ¡ .01).

Blaming the speaker We performed two-way ANOVA to analyze the effect of the speaker being NS/NNS and the performance of ASR on how listeners blame the speakers for the errors in the caption (Figure 6.4-right). We found that whether a speaker is NS/NNS has a statistically significant effect ($F_{1,418}$ = 112.07, p ¡ .001, η_p^2 = 0.21), while **performance of ASR did not have a significant effect on how much listeners blame the speaker** for the errors in the caption ($F_{2,417}$ = 1.32, not significant). We also did not observe a significant interaction $_{2,414}$ = 0.65, not significant).

Post-hoc test (ART-C contrasts with Tukey adjustment [200]) results showed that listeners **blamed the speaker significantly more when the speaker is NNS** (M = 4.90, SD = 1.52) compared to when the speaker is NS (M = 3.28, SD = 1.56, p ¡ .001).

6.3.4 Perception difference when performance disparity exists

When the ASR's performance was given higher when the speaker is NS than when the speaker is NNS (NS: WER = 5, NNS: WER = 30), we found that listeners perceived the ASR system to be of lower quality but found it more useful, while perceiving the speaker negatively when the speaker is NNS. Moreover, listeners blamed the speaker more and the ASR system less when the speaker was NNS. We report detailed results below.

RQ1: Perception towards ASR system and its output

Quality of ASR system We performed Mann-Whitney U test and found that **listeners perceived the quality of the ASR system significantly higher when the speaker was NS (med = 6) compared to NNS (med = 3), although the performance was lower in NNS's condition** (Figure 6.5-left) (U = 871, n_{NS} = n_{NNS} = 70, p ¡ .001, r = 0.64).

Usefulness of captions We performed Mann-Whitney U test and found **no significant difference exists on how listeners perceive the usefulness of captions between when the speaker is NS and when the speaker is NNS, although the performance was lower in NNS's condition** (Figure 6.5-right) (U = 2117, $n_{NS} = n_{NNS}$ = 70, p ¿ .05).

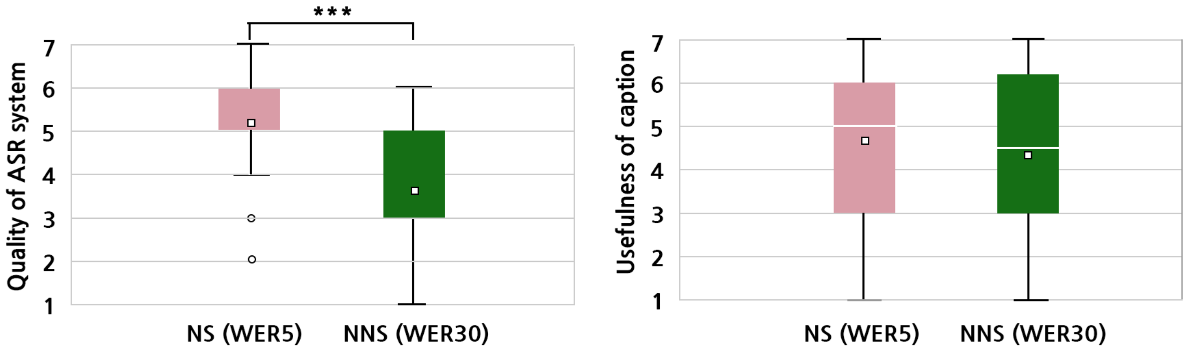


Figure 6.5: Listeners' perception towards the ASR system and its output given performance disparity between NS and NNS (*** p ¡ .001)

RQ2: Perception towards the speaker and their speech

We performed Mann-Whitney U test and found that listeners perceived significantly negatively when the speaker is NNS compared to when the speaker is NS when performance disparity exists (Figure 6.6). Listeners perceived the **negative effect of the speaker's accent on their understanding of the video content to be higher** (U = 385, med_{NS} = 1, med_{NNS} = 4, $n_{NS} = n_{NNS}$ = 70, p ¡ .001, r = 0.84), **expertise of the speaker to be lower** (U = 1620, med_{NS} = 6, med_{NNS} = 5, n_{NS} = 66, n_{NNS} = 63, p ¡ .05, r = 0.22), **reliability of the speaker's explanation to be lower** (U = 1409, med_{NS} = 6, med_{NNS} = 6, n_{NS} = 63, n_{NNS} = 59, p ¡ .05, r = 0.24), and **delivery skill of the speaker to be lower** (U = 995, med_{NS} = 6, med_{NNS} = 4, $n_{NS} = n_{NNS}$ = 70, p ¡ .001, r = 0.59) **when the speaker is NNS** compared to when the speaker is NS.

RQ3: Error blaming

We performed Mann-Whitney U test and found that listeners blamed the ASR system and the speaker significantly differently according to whether the speaker is NS/NNS even when performance disparity exists (Figure 6.7). Listeners blamed **the ASR system significantly less** (U = 1791, med_{NS} = 6, med_{NNS} = 4.5, $n_{NS} = n_{NNS}$ = 70, p ¡ .001, r = 0.27) and **the speaker significantly more** (U = 913, med_{NS} = 3, med_{NNS} = 5, $n_{NS} = n_{NNS}$ = 70, p ¡ .001, r = 0.63) **when the speaker is NNS** compared to when the speaker is NS.

6.4 Discussion

We discuss our interpretation of the results, design implications for building a more inclusive ASR model and its applications, and the generalizability of the results.

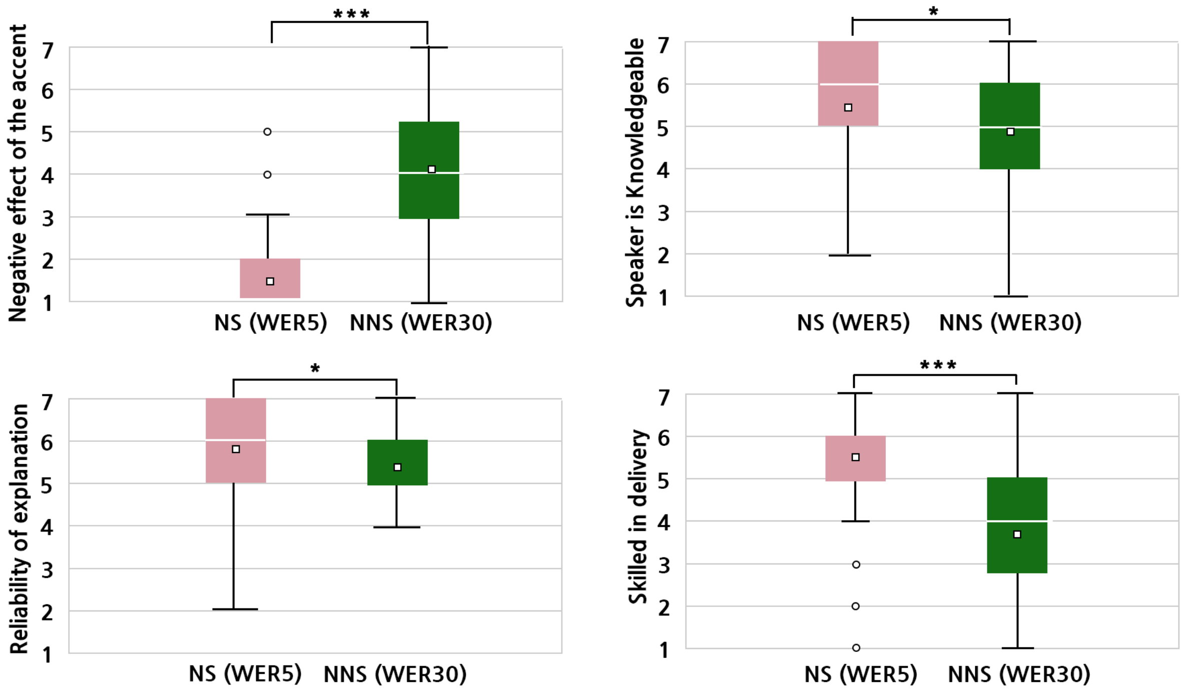


Figure 6.6: Listeners' perception towards the speaker and their speech given performance disparity between NS and NNS (* *p* ¡ .05, *** *p* ¡ .001)

6.4.1 Interpretation of Results

Here we discuss our interpretation of the results and their implications.

Performance Disparity of ASR System and Change of Its Usage

Our study has found that even with the same performance, ASR is more useful for understanding NNS's speech than NS's speech (Figure 6.2). In contrast, when the usage of ASR was first introduced in computer-mediated communications, many studies focused on how it can support NNS in understanding NS's speech [194, 223, 207, 229, 224, 242, 199, 222]. ASR systems were relatively not helpful for supporting the understanding of NNS's speech and even suggested removing NNS's captions as NNS was bothered by the low-quality captions compared to NS [205]. This could be because previously, ASR models were based on hidden Markov models (HMM) to convert audio to phoneme [238], so the performance of an ASR system for NNS could have been much lower. However, with the adoption of deep learning in ASR, there was a huge advance in the performance for various accents during the past decade [238], which may lead to a change in ASR's usage towards understanding NNS's speech. Moreover, since the room for increasing NNS's performance still remains significant (Section 6.3.1), we expect if this improvement is made in the future, this could also contribute to the change in ASR's usage.

To keep in line with this change in ASR's usage and make ASR technologies more useful, much research needs to be conducted. Previous studies have pointed out that WER 20% is a critical point for captions to be useful [224, 194]. However, they were investigated in the shoes of NNS understanding NS's speech. Our results suggest that the critical point of the captions' usefulness could be even different according to who the speaker is; the critical point of captions' usefulness for NNS's speech could be lower than that of NS's since our results suggest that the usefulness of captions of same performance differs according to whether the speaker is NS or NNS. Thus, investigating critical points of the captions'

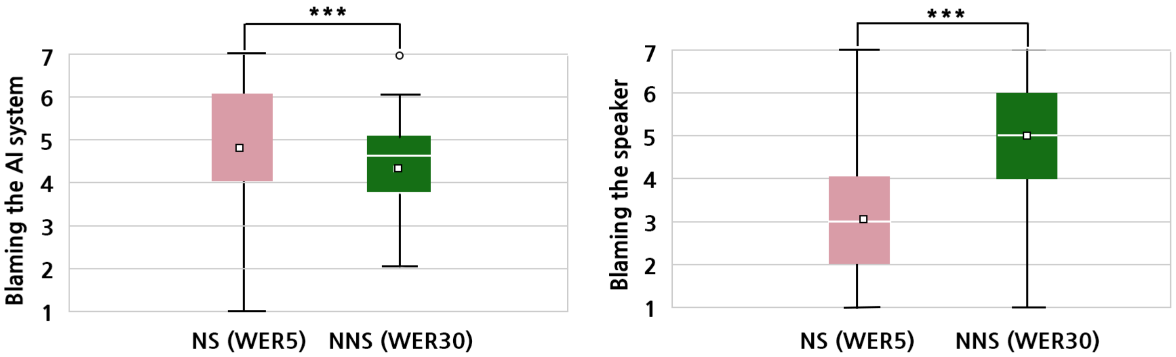


Figure 6.7: How much listeners blame errors in the captions on the AI system (left) and the speaker (right) given performance disparity between NS and NNS (*** p ¡ .001)

usefulness across various speakers is needed.

Furthermore, extensive research is needed to improve the UI/UX of ASR systems in alignment with the change in ASR's usage. Previous work focused more on ways to increase the usefulness of the ASR results of NS's speech, such as allowing NS to highlight important parts of their speech in the transcript [222] or allowing NS to edit their transcript [205]. However, the same UI/UX may not be applicable to NNS [205] as they face a higher cognitive load when communicating with others as they are relatively less familiar with the language [208]. Moreover, while NNS face more speaking anxiety than NS [186], their anxiety levels may further increase if they encounter errors in the captions of their own speech than NS. Thus, improving the UI/UX of ASR systems for NNS's speech should be further investigated to fully support the usefulness of ASR systems for NNS's speech.

Aggravating Unfavorable Situations for NNS

Even without ASR, previous work has suggested that NNS are more likely to be situated in an unfavorable situation or receive a more negative impression, such as being perceived as less credible compared to NS [213]. Our result also shows that NNS are perceived negatively (Figure 6.3). Now, with the use of ASR, our results suggest that in addition to receiving a more negative impression than NS, they also get more blame for errors in the captions than NS by using ASR technology for their speech (Section 6.3.4). Although ASR has the potential to support an understanding of NNS's speech, this suggests that the use of ASR technology can bring a more unfavorable environment for NNS by deteriorating the situation.

Furthermore, ASR models are now expanding their usage beyond supporting multilingual communications and being incorporated into technologies for other purposes, ranging from generating automatic transcripts in video-sharing platforms to intelligent personal assistants (IPA). NNS using those technologies may face difficulties as they are less comfortable with the language than NS in the first place. For instance, previous work found that NNS face more difficulties using IPA compared to NS as they have a hard time syntactically constructing commands in a foreign language [240]. On top of this unfavorable situation for NNS, the performance disparity of ASR may aggravate this situation; NNS may more heavily focus on their pronunciation when using IPA [241] due to the low performance, resulting in unequal user experience compared to NS. Likewise, although ASR itself may not be the main purpose of a certain technology, the performance disparity of ASR may significantly aggravate NNS's technology usage.

Likewise, Kentaro Toyama's Law of Amplification [233, 234] states that technology may widen the existing inequality unless there is an effort to reduce the inequality. Thus, to better accommodate ASR technology in various scenarios, it is crucial to be aware of the additional inequalities it could bring and design systems in a way that will decrease the unfair situation between the user groups.

Reason for Difference in Listeners' Perceptions Despite Identical ASR Performances

Our results suggest that although the performance of ASR is the same, listeners perceive the ASR result differently (Section 6.3.4). Previous work has suggested perceived accuracy may be different from the model accuracy [226]. However, in our study, the reason why the listeners' perception was different was not because the perceived accuracy was different; we found no significant difference in how they perceived the quality of the ASR system between NS and NNS (Section 6.3.3).

Instead, as our result (Section 6.3.2) suggests, this could be because people are more familiar with NS's accent, allowing them to hear better. Since NNS's speech was harder to hear, this could have led to higher usefulness of ASR output for NNS's speech even with the same ASR performance (Figure 6.2). Another potential reason behind blaming NNS more for the errors could be because people may have a subconscious bias that the NS's accent is 'better' than that of the NNS. Despite the effort of acknowledging diverse English accents as 'World Englishes' [209] in academia, bias on accents still persists in the world; previous studies have found that NNS aspired to match US or British accents and considered these accents to be the 'correct' accent [192, 216, 185].

This familiarity and bias towards certain accents may be shaped by education and media exposure. When learning English, students get frequent exposure to US or British accents in English language teaching materials [192]. English listening comprehension tests, such as TOEFL iBT, also only encompass accents of NS from certain countries (e.g., North America, the U.K., New Zealand, or Australia) [215, 202]. Furthermore, nationality is often treated as a proxy to qualify as a 'good' English teacher in the recruitment process [184]. Moreover, media may also play a significant role in one's familiarity towards certain accents and further shaping a subconscious bias. Research shows that in the media, 'non-standard' English speakers rarely appear compared to 'standard' English speakers, and even if they do so, are depicted as less favorable regarding their social or economic status and physical appearance [198]. On the other hand, news reporters are trained to speak in a certain accent [219]. Thus, to further accommodate English varieties and reduce a subconscious bias, societal measures should be taken for a gradual change in how people think and perceive different accents.

6.4.2 Design Implications

We present design implications for building inclusive ASR systems based on our study results.

Developing ASR Models

Despite various attempts to reduce the disparity gap in research [195, 237], similar to previous studies [196], our results also suggest that a huge performance disparity exists between NS and NNS (Section 6.3.1), while this disparity may form unfavorable situations for certain user groups. Therefore, researchers developing ASR models should put a much higher priority on reducing the disparity gap. In addition, our results show that the commercialized online platforms integrating ASR technology (e.g., YouTube, Zoom, Otter.ai) showed huge performance disparity compared to other models. This shows the needs for a quicker integration of state-of-the-art ASR models into commercialized services. Moreover, it is also important to explicitly address the performance disparity across diverse speaker groups in the reports or API documentations so that the platform builders could clearly be aware of the issue and select an appropriate model according to their purpose.

Furthermore, our results suggest that there still exist unfavorable situations for NNS regarding listeners' perceptions although model performance is the same (Section 6.3.3). Thus, for situations where listeners' perceptions are important, such as when ASR is used in job interviews, ASR models can be trained to incorporate listeners' perceptions, inspired by Reinforcement Learning from Human Feedback (RLHF) [246, 231, 221]. We first construct a speech and ASR result dataset with listeners' perceptions (e.g., level of blaming speakers) annotated. Then, we can train a model to predict listeners' perceptions. Based on the prediction model, we integrate the difference in listeners' perceptions between NNS and NS as the reward signal in the ASR model, reducing the difference. This is different from RLHF, as our suggested reward allows the ASR model to learn to have similar listeners' perceptions towards different speaker groups. Although this may result in performance trade-offs, a fairer ASR model may be more desirable than one with just a higher average performance depending on one's purpose.

Systems Utilizing ASR Models

With the fast-evolving AI technologies and their rapid development, there exist various ASR models with different performances. Hence, when building systems that utilize ASR models, selecting an appropriate ASR model while considering its purpose would be important. Previously, this selection of the ASR model could have been solely dependent on model performance, but our results suggest that listeners' perception should be another important factor to consider. The choice of the ASR model can even depend more on differences in listeners' perception in circumstances where it is important to prevent yielding any discriminating or unfair decisions. For instance, when building an ASR system to support job interviews, preventing speakers from receiving an unfair impression while using the system should be a more important factor to consider when selecting which ASR model to incorporate compared to casual conversation settings. Selecting the ASR model focusing only on its high performance neglecting listeners' different perceptions can cause group inequity between NS and NNS. Thus, system developers should take a step back and consider the societal impact the system they are building would bring as their system would be used by various users and may shape unconscious bias.

Furthermore, speaker-adaptive ASR system is needed where different UI/UX could be provided based on who the speaker is. Previously, listener-adaptive subtitling systems have been proposed, where the rendering of subtitles is personalized to the listener and their device environment [187]. We suggest that UI/UX personalization also needs to be considered in the perspective of speakers to mitigate the unfair environment for NNS when designing an ASR system. For instance, considering our results that users tend to blame the NNS more, when NNS is speaking, the ASR system could explicitly indicate that the model is not performing well so that the listeners could be aware of AI's fault instead of focusing on NNS's pronunciation.

6.4.3 Generalization of Results

Our study focused on investigating how listeners perceive ASR results differently according to the speaker's accent and how it can result in unfavorable situations for NNS. This result may generalize to other AI systems where users may make value judgments on the input itself or the person who generates the input. For instance, for the output of a grammatical error correction model showing similar accuracy and errors for the same lines of writing of NS and NNS, readers may attribute the errors differently as they may have a prejudice that NS is better at using the language.

This biased perception difference not only occurs between NS and NNS. For example, when a handwriting recognition model fails to recognize certain handwriting, users may perceive differently based on the value judgment of how much the handwriting is aesthetically appealing. This may result in more severe consequences if it is related to diverse accessibility issues. For instance, if the gesture recognition model fails to recognize someone with motor impairments, people may blame the person for the errors, which may induce unjust outcomes. Thus, our results and design implications could be applied to other AI systems as well.

6.5 Limitation & Future Work

We acknowledge the limitations of our study and present possible future work.

First, although we divided speakers into NS and NNS, the way of speaking can vary greatly within each group. People from the same country can have different accents depending on their socioeconomic and sociolinguistic factors [235]. In addition, listeners may perceive ASR results differently for NS from countries where they use multiple languages as their official language (e.g., India, Kenya, the Philippines). Moreover, there exist other factors of speech that may influence the results that we did not take into consideration: the speaker's age, gender, voice tone, and speech fluency and speed could affect the listener's overall experience and perception of the speaker [210, 206]. Although focusing on NS and NNS's accents could be a meaningful start in understanding differences in listeners' perceptions, future studies could include more speakers considering various factors and investigating how these factors play a role in how listeners perceive.

Second, the listener's familiarity and personal preference towards how the speaker speaks may also affect how the listener perceives the ASR result. Since we found that the people are more likely to be familiar with NS's accent, while not being familiar with NNS's accents (Section 6.3.2), we assigned listeners who are non-familiar with speaker's accents for NNS, while assigning listeners who are familiar with speaker's accents for NS in our study. However, further study is needed on how the listener's perception differs in other combinations, such as being familiar with NNS's accents.

Lastly, our experiments mainly focused on a single scenario (i.e., a listener watching a video of a speaker asynchronously with auto-generated captions), which may differ from other scenarios using ASR systems. For instance, listeners may perceive differently if they watch a speaker synchronously. In addition, the listener's perception could be influenced by the listeners listening together; the composition of other listeners (e.g., other listeners being the same ethnicity as the speaker) or reactions of other listeners (e.g., applause) could impact their perceptions. Thus, future work could investigate other scenarios for a more generalizable result.

Chapter 7. Discussion

This thesis found the additional risks faced by user groups who are overlooked during the design process of the AI system when they interact with AI systems by understanding the (1) usage pattern differences in video recommendation systems, (2) usage pattern differences in large language models, (3) perception differences in behavior log-based personality detection systems, and (4) perception differences in automatic speech recognition systems.

This section discusses (1) how to identify disadvantaged users in AI systems and (2) strategies for reducing these disadvantages: informing users and society about the disadvantages, and building inclusive AI systems.

7.1 Identifying Disadvantaged Users in AI Systems

Understanding which user groups may be disadvantaged is a complex process, as no single group is universally disadvantaged across all AI systems. For instance, non-native speakers face challenges when using automatic speech recognition (ASR) systems or large language models (LLMs) for writing, but they might not encounter disadvantages in other AI applications. The key to identifying disadvantaged users is via analyzing the usage patterns or perceptions of the AI system and by revealing differences in usage patterns or perceptions from that of the main target user groups who are considered in the design process. For example, older adults being accustomed to watching traditional television may have influenced their interaction patterns to be less interactive than non-older adults in online video platforms. Which in turn may lead older adults to keep watching the video although the video platform is designed to recommend more videos by relying on users' usage patterns to drop off or skim the video if they are not interested in the video. This identification of the user groups and the challenges they face is the first step toward achieving inclusive AI systems.

7.2 Strategies for reducing the disadvantages users face

We can reduce the disadvantages the overlooked users face via (1) informing users and society about the disadvantages and (2) building inclusive AI systems.

7.2.1 Informing Users and Society About the Disadvantages

One way to reduce the disadvantages faced by overlooked users is to inform both the users themselves and society about the additional risks they encounter when interacting with AI systems. Awareness can be a powerful tool in mitigating these risks, as it enables both behavioral changes and systemic support.

First, it is essential to ensure that the users themselves are aware of the risks they may face. For example, privacy-sensitive users interacting with behavior-based personality detection systems may alter their behavior due to privacy concerns, leading to inaccurate results. If these users are unaware of how their privacy concerns influence the system's performance, they may mistakenly assume that their results are consistent with the overall system accuracy. By understanding that their specific concerns may impact the outcomes, these users can better manage their interactions with the system or adjust their expectations accordingly.

Similarly, non-native speakers (NNS) using large language models (LLMs) for writing tasks often rely heavily on AI-generated drafts, which can diminish the authenticity of their writing. If these users are informed that their writing may be inauthentic due to more likeliness to ask LLM for drafts or over-reliance on LLM-generated drafts, they may adjust their behavior by brainstorming or planning before requesting drafts. Such awareness can help mitigate the risks posed by their usage patterns, offering a temporary but practical solution.

In addition, informing society about the risks faced by overlooked users can encourage broader, systemic efforts to address these issues. For instance, societal perceptions towards non-native speakers' English may influence how non-native speakers are falsely blamed when using ASR systems. Errors in ASR output are often attributed to the speaker rather than the system when the speaker is a non-native English speaker. By raising awareness of this bias, society can work towards gradually changing negative perceptions towards non-native speakers' accents and fostering a more supportive environment for non-native speakers.

By informing the society about the risks, we can have society-level effort in reducing the risks or supporting those users. For instance, since non-native speakers' writing patterns of directly starting to write without brainstorming may lead them to directly ask for drafts to LLM for writing without brainstorming, for example in a writing class, instructors may want to additionally instruct non-native speakers on how to utilize LLMs for brainstorming and planning before the writing.

7.2.2 Building inclusive AI systems

Furthermore, by building inclusive AI systems, we can reduce the additional risk in the future. First, the guiding principles that shape the development of AI systems should be reconsidered that relates to the disadvantages faced by overlooked user groups. While many design principles are intended to guide AIs to provide benefits to the users, as demonstrated throughout this thesis, these principles can inadvertently impose risks on users whose usage patterns or perceptions deviate from those of the majority users. Therefore, revising these principles is essential to ensure that future AI systems do not further impose risks to certain users.

Another way to build inclusive AI systems is through adaptive design approaches that account for the diverse usage patterns and perceptions of different user groups. Rather than relying on a one-size-fits-all model, AI systems can be designed to dynamically adjust based on a user's usage patterns or perceptions. For example, rather than relying on users' prompt and passively respond to those prompts, writing assisting systems could actively track users writing patterns. If the user used not much time in planning or brainstorming and directly asks for the draft, the LLM could rather focus more on showing short drafts of different ideas rather than giving one full draft that non-native speakers may easily copy-paste. By embedding adaptability into system designs, AI technologies can better align with the diverse experiences of their users, ensuring equitable access and benefits for all.

Chapter 8. Conclusion

In conclusion, this thesis provides critical insights into how AI systems can disadvantage user groups overlooked during the design process and has different usage patterns or perceptions of AI systems from that of main target users considered during the development of the system. This thesis investigated this in diverse AI systems, i.e., video recommendation systems, large language models, behavior log-based personality detection systems, and automatic speech recognition systems. By revealing the disadvantages certain users face followed by fostering the awareness of these disadvantages and rethinking design principles for the future AI development, we can expect AI systems to be more inclusive that benefit all users, regardless of their usage patterns or perceptions.

Bibliography

[1] Angel Hsing-Chi Hwang, Q Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler. 2024. " It was 80% me, 20% AI": Seeking Authenticity in Co-Writing with Large Language Models. *arXiv preprint arXiv:2411.13032* (2024).

[2] Vinzenz Wolf and Christian Maier. 2024. ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management* 79 (2024), 102821.

[3] Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query Refinement Prompts for Closed-Book Long-Form QA. In *Annual Meeting of the Association for Computational Linguistics.*

[4] Khaled Barkaoui. 2019. WHAT CAN L2 WRITERS'PAUSING BEHAVIOR TELL US ABOUT THEIR L2 WRITING PROCESSES? *Studies in Second Language Acquisition* 41, 3 (2019), 529–554.

[5] Jessie S Barrot. 2023. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing* 57 (2023), 100745.

[6] Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, Paula Buttery, and Mark Aronoff. 2015. Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLOS ONE* 10 (2015).

[7] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[8] Yuexi Chen and Zhicheng Liu. 2024. WordDecipher: Enhancing Digital Workspace Communication with Explainable AI for Non-native English Speakers. *ArXiv* abs/2404.07005 (2024).

[9] Yuh-show Cheng. 2002. Factors associated with foreign language writing anxiety. *Foreign language annals* 35, 6 (2002), 647–656.

[10] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *ArXiv* abs/2107.07430 (2021).

[11] Julio Roca De Larios, Rosa Manchón, Liz Murphy, and Javier Marín. 2008. The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing* 17, 1 (2008), 30–47.

[12] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Tak Yeon Lee, So-Yeon Ahn, Alice Oh, and Acknowledgment Negotiation Answer. 2023. Exploring Student-ChatGPT Dialogue in EFL Writing Education. In *Thirty-seventh Conference on Neural Information Processing Systems, Neural information processing systems foundation.*

[13] Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jacky Jamieson, and Kentaro Inui. 2023. Use of an AI-powered Rewriting Support Software in Context

with Other Tools: A Study of Non-Native English Speakers. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023).

[14] Yateendra Joshi. 2013. Why does English dominate science publishing. *Editage Insights* (2013).

[15] Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models. In *Conference on Empirical Methods in Natural Language Processing.*

[16] Yewon Kim, Thanh-Long V Le, Donghwi Kim, Mina Lee, and Sung-Ju Lee. 2024. How Non-native English Speakers Use, Assess, and Select AI-Generated Paraphrases with Information Aids. *arXiv preprint arXiv:2405.07475* (2024).

[17] Ariyanti Ariyanti and Rinda Fitriana. 2017. EFL students' difficulties and needs in essay writing. In *International Conference on Teacher Training and Education 2017 (ICTTE 2017).* Atlantis Press, 32–42.

[18] Sigrun Biesenbach-Lucas. 2007. Students writing emails to faculty: An examination of e-politeness among native and non-native speakers of English. (2007).

[19] Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2024. Native vs Non-Native Language Prompting: A Comparative Analysis. *ArXiv* abs/2409.07054 (2024).

[20] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *ArXiv* abs/2304.05613 (2023).

[21] Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking Clarification Questions to Handle Ambiguity in Open-Domain QA. *ArXiv* abs/2305.13808 (2023).

[22] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[23] Yoonsang Lee, Xi Ye, and Eunsol Choi. 2024. AmbigDocs: Reasoning across Documents on Different Entities under the Same Name. In *First Conference on Language Modeling.*

[24] Grace Li, Tao Long, and Lydia B. Chilton. 2023. Eliciting Topic Hierarchies from Large Language Models. *ArXiv* abs/2310.19275 (2023).

[25] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–25.

[26] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Y. Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4 (2023).

[27] Stephen Macneil, Andrew Tran, Joanne Kim, Ziheng Huang, Seth Bernstein, and Dan Mogil. 2023. Prompt Middleware: Mapping Prompts for Large Language Models to UI Affordances. *ArXiv* abs/2307.01142 (2023).

[28] Ali Malik, Stephen Mayhew, Chris Piech, and K. Bicknell. 2024. From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation. *ArXiv* abs/2406.03030 (2024).

[29] Bertalan Meskó. 2023. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *Journal of Medical Internet Research* 25 (2023).

[30] Kristyan Spelman Miller, Eva Lindgren, and Kirk PH Sullivan. 2008. The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *Tesol Quarterly* 42, 3 (2008), 433–454.

[31] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Conference on Empirical Methods in Natural Language Processing*.

[32] Timotius Pradana A. Moelyono, Elisabet Titik Murtisari, Daniel Kurniawan, and Andrew Thren. 2023. Google Translate in EFL Freshmen's Writing Assignments: Uses, Awareness of Benefits and Drawbacks, and Perceived Reliance. *Vision: Journal for Language and Foreign Language Learning* (2023).

[33] Ali Hakimi Parizi, Yuyang Liu, Prudhvi Nokku, Sina Gholamian, and David B. Emerson. 2023. A Comparative Study of Prompting Strategies for Legal Text Classification. *Proceedings of the Natural Legal Language Processing Workshop 2023* (2023).

[34] Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users.

[35] Manon Reusens, Philipp Borchert, Jochen De Weerdt, and Bart Baesens. 2024. Native Design Bias: Studying the Impact of English Nativeness on Language Model Performance.

[36] Andrea Révész, Marije Michel, Xiaojun Lu, Nektaria Kourtali, Minjin Lee, and Laís Borges. 2022. The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing* 58 (2022), 100927.

[37] Mohammod Moninoor Roshid, Susan Webb, and Raqib Chowdhury. 2018. English as a Business Lingua Franca: A Discursive Analysis of Business E-Mails. *International Journal of Business Communication* 59 (2018), 83 – 103.

[38] Miyuki Sasaki. 2002. Building an empirically-based model of EFL learners' writing processes. *New directions for research in L2 writing* (2002), 49–80.

[39] Orit Shaer, Angel Cooper, Osnat Mokryn, Andrew L. Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024).

[40] Shuming Shi, Enbo Zhao, Duyu Tang, Yan Wang, Piji Li, Wei Bi, Haiyun Jiang, Guoping Huang, Leyang Cui, Xinting Huang, Cong Zhou, Yong Dai, and Dongyang Ma. 2022. Effidit: Your AI Writing Assistant. *ArXiv* abs/2208.01815 (2022).

[41] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[42] Daphne van Weijen Dr. 2012. The language of (future) scientific communication. *Research trends* 1, 31 (2012), 3.

[43] Qian Wan, Si-Yuan Hu, Yu Zhang, Pi-Hui Wang, Bo Wen, and Zhicong Lu. 2023. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction* 8 (2023), 1 – 26.

[44] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, Chenxi Yue, Haiyang Zhang, Yi-Hsueh Liu, Xiang Li, Bao Ge, Dajiang Zhu, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. 2023. Prompt Engineering for Healthcare: Methodologies and Applications. *ArXiv* abs/2304.14670 (2023).

[45] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[46] Michael J.Q. Zhang and Eunsol Choi. 2023. Clarify When Necessary: Resolving Ambiguity Through Interaction with LMs. *ArXiv* abs/2311.09469 (2023).

[47] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *ArXiv* abs/2304.04675 (2023).

[48] Mark S Allen, Emma E Walter, and Máirtín S McDermott. 2017. Personality and sedentary behavior: A systematic review and meta-analysis. *Health Psychology* 36, 3 (2017), 255.

[49] Lyndsey L Bakewell, Konstantina Vasileiou, Kiel S Long, Mark Atkinson, Helen Rice, Manuela Barreto, Julie Barnett, Michael Wilson, Shaun Lawson, and John Vines. 2018. Everything we do, everything we press: Data-driven remote performance management in a mobile workplace. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 371.

[50] Gene Ball and Jack Breese. 2000. Relating personality and behavior: posture and gestures. In *Affective interactions*. Springer, 196–203.

[51] Kirstie Ball. 2010. Workplace surveillance: An overview. *Labor History* 51, 1 (2010), 87–106.

[52] Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology* 44, 1 (1991), 1–26.

[53] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. 2011. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 255–262.

[54] Shlomo Berkovsky, Ronnie Taib, Irena Koprinska, Eileen Wang, Yucheng Zeng, Jingjie Li, and Sabina Kleitman. 2019. Detecting Personality Traits Using Eye-Tracking Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 221.

[55] Tim Blumer and Nicola Döring. 2012. Are we the same online? The expression of the five factor personality traits on the computer and the Internet. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 6, 3 (2012).

[56] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[57] Neil Christiansen and Robert Tett. 2013. *Handbook of personality at work.* Routledge.

[58] Philip J Corr and Gerald Matthews. 2009. *The Cambridge handbook of personality psychology.* Cambridge University Press Cambridge, UK:.

[59] CPP. 2018. CPP — The MyersBriggs®Company. (2018). https://www.cpp.com/ Last Accessed: 2018-08-09.

[60] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. 2013. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction.* Springer, 48–55.

[61] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[62] Alain Dössegger, Nicole Ruch, Gerda Jimmy, Charlotte Braun-Fahrländer, Urs Mäder, Johanna Hänggi, Heidi Hofmann, Jardena J Puder, Susi Kriemler, and Bettina Bringolf-Isler. 2014. Reactivity to accelerometer measurement of children and adolescents. *Medicine and science in sports and exercise* 46, 6 (2014), 1140.

[63] John T Foley, Michael W Beets, and Bradley J Cardinal. 2011. Monitoring children's physical activity with pedometers: Reactivity revisited. *Journal of Exercise Science & Fitness* 9, 2 (2011), 82–86.

[64] Howard S Friedman, M Robin DiMatteo, and Angelo Taranta. 1980. A study of the relationship between individual differences in nonverbal expressiveness and factors of personality and social interaction. *Journal of Research in Personality* 14, 3 (1980), 351–364.

[65] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems.* ACM, 253–262.

[66] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[67] Liang Gou, Michelle X Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 955–964.

[68] Heather Hausenblas and Ryan E Rhodes. 2016. ExErcisE Psychology. (2016).

[69] Paul P Heppner, Bruce E Wampold, and Dennis M Kivlighan. 2007. Research design in counseling: Research, statistics, & program evaluation. *Cengage Learning* (2007).

[70] Everard Ho and Vichita Vathanophas. 2003. Relating personality traits and prior knowledge to focus group process and outcome: an exploratory research. *PACIS 2003 Proceedings* (2003), 67.

[71] Alan E Kazdin. 1974. Reactive self-monitoring: the effects of response desirability, goal setting, and feedback. *Journal of consulting and clinical psychology* 42, 5 (1974), 704.

[72] Alan E Kazdin. 1979. Unobtrusive measures in behavioral assessment. *Journal of Applied Behavior Analysis* 12, 4 (1979), 713–724.

[73] Seoyoung Kim, Jiyoun Ha, and Juho Kim. 2018. Detecting Personality Unobtrusively from Users' Online and Offline Workplace Behaviors. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW515.

[74] Sandjar Kozubaev, Fernando Rochaix, Carl DiSalvo, and Christopher A Le Dantec. 2019. Spaces and Traces: Implications of Smart Technology in Public Housing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 439.

[75] Kibeom Lee, Michael C Ashton, and Kang-Hyun Shin. 2005. Personality correlates of workplace anti-social behavior. *Applied Psychology* 54, 1 (2005), 81–98.

[76] Weijian Li, Yuxiao Chen, Tianran Hu, and Jiebo Luo. 2018. Mining the Relationship between Emoji Usage Patterns and Personality. *arXiv preprint arXiv:1804.05143* (2018).

[77] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[78] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P Dow. 2016. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 260–273.

[79] Jennifer Dodorico McDonald. 2008. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire* 1, 1 (2008), 1–19.

[80] Walter Mischel. 2013. *Personality and assessment*. Psychology Press.

[81] Isabel Briggs Myers, Mary H McCaulley, and Robert Most. 1985. *Manual, a guide to the development and use of the Myers-Briggs type indicator*. consulting psychologists press.

[82] Daniel Olguın Olguın, Peter A Gloor, and Alex Sandy Pentland. 2009. Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, Vol. 9.

[83] Chanda Phelan, Cliff Lampe, and Paul Resnick. 2016. It's creepy, but it doesn't bother me. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 5240–5251.

[84] Brent W Roberts. 2009. Back to the future: Personality and assessment and personality development. *Journal of research in personality* 43, 2 (2009), 137–145.

[85] Ivan Robertson and Militza Callinan. 1998. Personality and work behaviour. *European Journal of Work and Organizational Psychology* 7, 3 (1998), 321–340.

[86] Christine Satchell and Paul Dourish. 2009. Beyond the user: use and non-use in HCI. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*. ACM, 9–16.

[87] Stefan Schneegass, Romina Poguntke, and Tonja Machulla. 2019. Understanding the Impact of Information Representation on Willingness to Share Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 523.

[88] Lee Taber and Steve Whittaker. 2018. Personality depends on the medium: differences in self-perception on Snapchat, Facebook and offline. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 607.

[89] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A Smith. 2015. Can an Algorithm Know the Real You?: Understanding People's Reactions to Hyper-personal Analytics Systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 797–806.

[90] Ziang Xiao, Michelle X Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 437–447.

[91] Lingling Xu, Cheng Yi, and Yunjie Xu. 2007. Emotional expression online: The impact of task, relationship and personality perception on emoticon usage in instant messenger. *PACIS 2007 Proceedings* (2007), 79.

[92] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. 2019. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662* (2019).

[93] Academic Credit Bank System (ACBS). 2022. https://www.cb.or.kr/creditbank/info/nInfo7_1.do

[94] Yakup Akgül et al. 2018. Web accessibility of MOOCs for elderly students: The case of Turkey. *Journal of Life Economics* 5, 4 (2018), 141–150. DOI:https://doi.org/10.15637/jlecon.266

[95] Subhashni Appana. 2008. A review of benefits and limitations of online learning in the context of the student, the instructor and the tenured faculty. *International Journal on E-learning* 7, 1 (2008), 5–22. https://www.learntechlib.org/primary/p/22909/

[96] David Aspin and Judith Chapman. 2001. Lifelong learning: concepts, theories and values. In *Proceedings of the 31st Annual Conference of SCUTREA*. University of East London: SCUTREA, 38–41. DOI:https://doi.org/10.1080/026013700293421

[97] Xue Bai, Yiqin He, and Florian Kohlbacher. 2020. Older people's adoption of e-learning services: a qualitative study of facilitators and barriers. *Gerontology & geriatrics education* 41, 3 (2020), 291–307.

[98] Paul B Baltes and Margret M Baltes. 1990. Psychological perspectives on successful aging: The model of selective optimization with compensation. (1990).

[99] Sharon Jeffcoat Bartley and Jennifer H Golek. 2004. Evaluating the cost effectiveness of online and face-to-face instruction. *Journal of Educational Technology & Society* 7, 4 (2004), 167–175.

[100] Paul Bélanger. 2015. *Self-construction and social transformation: Lifelong, lifewide and life-deep learning*. UNESCO Institute for Lifelong Learning.

[101] Paola Beltran, Paul Rodriguez-Ch, and Priscila Cedillo. 2017. A Systematic Literature Review for Development, Implementation and Deployment of MOOCs Focused on Older People. In *2017 International Conference on Information Systems and Computer Science (INCISCOS)*. IEEE, 287–294.

[102] Frank Bentley and Janet Murray. 2016. Understanding video rewatching experiences. In *Proceedings of the ACM international conference on interactive experiences for TV and online video*. 69–75. DOI: https://doi.org/10.1145/2932206.2932213

[103] Frank Bentley, Max Silverman, and Melissa Bica. 2019. Exploring online video watching behaviors. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. 108–117. DOI:https://doi.org/10.1145/3317697.3323355

[104] Nicolas Biard, Salomé Cojean, and Eric Jamet. 2018. Effects of segmentation and pacing on procedural learning by video. *Computers in Human Behavior* 89 (2018), 411–417. DOI:https://doi.org/10.1016/j.chb.2017.12.002

[105] J Martin Bland and Douglas G Altman. 2000. The odds ratio. *Bmj* 320, 7247 (2000), 1468. DOI: https://doi.org/10.1136/bmj.320.7247.1468

[106] Way Kiat Bong and Weiqin Chen. 2016. How accessible are MOOCs to the elderly?. In *International Conference on Computers Helping People with Special Needs*. Springer, 437–444. DOI:https://doi.org/10.1007/978-3-319-41264-1_60

[107] Roger Boshier. 1977. Motivational orientations re-visited: Life-space motives and the education participation scale. *Adult education* 27, 2 (1977), 89–115. DOI:https://doi.org/10.1177/074171367702700202

[108] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012). DOI:https://doi.org/10.1037/13620-004

[109] Christopher G Brinton, Swapna Buccapatnam, Mung Chiang, and HV Poor. 2015. Mining MOOC clickstreams: On the relationship between learner behavior and performance. *arXiv preprint arXiv:1503.06489* (2015).

[110] Katherine Brookfield, Sara Tilley, and Máire Cox. 2016. Informal science learning for older adults. *Science Communication* 38, 5 (2016), 655–665.

[111] Christopher Brooks, Joshua Gardner, and Kaifeng Chen. 2018. How gender cues in educational video impact participation and retention. International Society of the Learning Sciences, Inc.[ISLS].

[112] Isaac Chuang and Andrew Ho. 2016. HarvardX and MITx: Four years of open online courses–fall 2012-summer 2016. *Available at SSRN 2889436* (2016). DOI:https://doi.org/10.2139/ssrn.2889436

[113] John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research.* Sage publications.

[114] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*. 293–296.

[115] Thomas S Dee. 2005. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review* 95, 2 (2005), 158–165. DOI:https://doi.org/10.1257/000282805774670446

[116] Mª del Pilar Díaz-López, Remedios López-Liria, José M Aguilar-Parra, and David Padilla-Góngora. 2016. Keys to active ageing: new communication technologies and lifelong learning. *SpringerPlus* 5, 1 (2016), 768. DOI:https://doi.org/10.1186/s40064-016-2434-8

[117] Ione Y DeOllos and David C Morris. 1999. The Internet as an information resource for older adults. *Journal of Educational Technology Systems* 28, 2 (1999), 107–120.

[118] Richard Desjardins, Marcella Milana, and Kjell Rubenson. 2006. *Unequal chances to participate in adult learning: International perspectives.* Number 83. Richard Desjardins.

[119] Michelle Dorin. 2007. Online education of older adults and its relation to life satisfaction. *Educational Gerontology* 33, 2 (2007), 127–143.

[120] Richard Dorsett, Silvia Lui, and Martin Weale. 2010. *Economic benefits of lifelong learning.* Centre for Learning and Life Chances in Knowledge Economies and Societies.

[121] Janet E. Truluck, Bradley C. Courtenay. 1999. Learning style preferences among older adults. *Educational gerontology* 25, 3 (1999), 221–236. DOI:https://doi.org/10.1080/036012799267846

[122] Maureen Ebben and Julien S Murphy. 2014. Unpacking MOOC scholarly discourse: A review of nascent MOOC scholarship. *Learning, media and technology* 39, 3 (2014), 328–345. DOI:https://doi.org/10.1080/17439884.2013.878352

[123] Kenneth F Ferraro and Janet M Wilmoth. 2013. *Gerontology: Perspectives and issues.* Springer Publishing Company.

[124] Brian Findsen and Marvin Formosa. 2011. *Lifelong learning in later life: A handbook on older adult learning.* Brill Sense.

[125] Jens Friebe and Bernhard Schmidt-Hertha. 2013. Activities and barriers to education for elderly people. *Journal of Contemporary Educational Studies/Sodobna Pedagogika* 64, 1 (2013).

[126] Ernest Furchtgott and Jerome R Busemeyer. 1981. Age preferences for professional helpers. *Journal of gerontology* 36, 1 (1981), 90–92. DOI:https://doi.org/10.1093/geronj/36.1.90

[127] Rod P Githens. 2007. Older adults and e-learning: Opportunities and barriers. *Quarterly Review of Distance Education* 8, 4 (2007), 329.

[128] Joselyn Goopio and Catherine Cheung. 2021. The MOOC dropout phenomenon and retention strategies. *Journal of Teaching in Travel & Tourism* 21, 2 (2021), 177–197.

[129] Mackenzie Robinson Graves. 2018. Lifelong learning: Applying cognitive load theory to elder learners suffering from age-related cognitive decline. *SFU Educational Review* 11, 1 (2018).

[130] Philip J Guo. 2017. Older adults learning computer programming: motivations, frustrations, and design opportunities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 7070–7083. DOI:https://doi.org/10.1145/3025453.3025945

[131] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 41–50. DOI:https://doi.org/10.1145/2556325.2566239

[132] Angela H Gutchess, Carolyn Yoon, Ting Luo, Fred Feinberg, Trey Hedden, Qicheng Jing, Richard E Nisbett, and Denise C Park. 2006. Categorical organization in free recall across culture and age. *Gerontology* 52, 5 (2006), 314–323.

[133] James W Hardin. 2005. Generalized estimating equations (GEE). *Encyclopedia of statistics in behavioral science* (2005).

[134] Luqman Hidayat, G Gunarhadi, and Furqon Hidayatulloh. 2017. Multimedia based learning materials for deaf students. *European Journal of Special Education Research* (2017).

[135] Laura Holyoke and Erick Larson. 2009. Engaging the adult learner generational mix. *Journal of Adult Education* 38, 1 (2009), 12–21.

[136] Jonas Langset Hustad, Andreas Schille, and Eirik Wattengård. 2019. Escaping the talking head: Experiences with three different styles of MOOC video. In *Proceedings of the the 6th European Conference on Massive Open Online Courses*. 151–156.

[137] The Chosun Ilbo. 2018. Workers Face Earlier Retirement Than Expected. http://english.chosun.com/site/data/html_dir/2018/11/01/2018110100695.html

[138] Consumer Insights. 2015. Attention spans.

[139] Bora Jin, Junghwan Kim, and Lisa M Baumgartner. 2019. Informal learning of older adults in using mobile devices: A review of the literature. *Adult Education Quarterly* 69, 2 (2019), 120–141. DOI:https://doi.org/10.1177/0741713619834726

[140] K-MOOC. 2021. http://www.kmooc.kr

[141] Juho Kim, Philip J Guo, Carrie J Cai, Shang-Wen Li, Krzysztof Z Gajos, and Robert C Miller. 2014a. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 563–572. DOI:https://doi.org/10.1145/2642918.2647389

[142] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014b. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. 31–40. DOI:https://doi.org/10.1145/2556325.2566239

[143] Jeongyeon Kim and Juho Kim. 2021. FitVid: Towards Development of Responsive and Fluid Video Content Adaptation. In *Workshop on Imagining Post-COVID Education with AI.*

[144] Jessica Kriegel. 2013. *Differences in learning preferences by generational cohort: Implications for instructional design in corporate web-based learning.* Drexel University.

[145] Marjan Laal. 2011a. Barriers to lifelong learning. *Procedia-Social and Behavioral Sciences* 28 (2011), 612–615. DOI:https://doi.org/10.1016/j.sbspro.2011.11.116

[146] Marjan Laal. 2011b. Lifelong learning: What does it mean? *Procedia-Social and Behavioral Sciences* 28 (2011), 470–474.

[147] Youngju Lee and Jaeho Choi. 2011. A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development* 59, 5 (2011), 593–618. DOI:https://doi.org/10.1007/s11423-010-9177-y

[148] Nan Li, Łukasz Kidziński, Patrick Jermann, and Pierre Dillenbourg. 2015. MOOC video interaction patterns: What do they tell us?. In *European Conference on Technology Enhanced Learning*. Springer, 197–210.

[149] Xingyu Liu, Patrick Carrington, Xiang'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[150] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation* (2005). DOI:https://doi.org/10.1108/00220410510632040

[151] David W Livingstone. 2001. Adults' informal learning: Definitions, findings, gaps and future research. (2001).

[152] Tharindu Rekha Liyanagunawardena and Shirley Ann Williams. 2016. Elderly learners and massive open online courses: a review. *Interactive journal of medical research* 5, 1 (2016), e4937. DOI:https://doi.org/10.2196/ijmr.4937

[153] Kate Manuel. 2002. Teaching information literacy to generation. *Journal of library administration* 36, 1-2 (2002), 195–217. DOI:https://doi.org/10.1300/J111v36n01_12

[154] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.

[155] Mary C Milliken, Susan O'Donnell, Kerri Gibson, and Betty Daniels. 2012. Older citizens and video communications: A case study. *The Journal of Community Informatics* 8, 1 (2012).

[156] National Academies of Sciences, Engineering, and Medicine and others. 2018. *How people learn II: Learners, contexts, and cultures*. National Academies Press.

[157] Andrew Ng and Jennifer Widom. 2014. Origins of the modern MOOC (xMOOC). *Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report* (2014), 34–47.

[158] Tuan Nguyen. 2015. The effectiveness of online learning: Beyond no significant difference and future horizons. *MERLOT Journal of Online Learning and Teaching* 11, 2 (2015), 309–319.

[159] Anna Nishchyk, Norun Christine Sanderson, Weiqin Chen, et al. 2017. How elderly people experience videos in MOOCs. In *DS 88: Proceedings of the 19th International Conference on Engineering and Product Design Education (E&PDE17), Building Community: Design Education for a Sustainable Future, Oslo, Norway, 7 & 8 September 2017*. 686–691.

[160] Ozlem Ozan and Yasin Ozarslan. 2016. Video lecture watching behaviors of learners in online courses. *Educational Media International* 53, 1 (2016), 27–41. DOI:https://doi.org/10.1080/09523987.2016.1189255

[161] Fred Paas, Gino Camp, and Remy Rikers. 2001. Instructional compensation for age-related cognitive declines: Effects of goal specificity in maze learning. *Journal of educational psychology* 93, 1 (2001), 181.

[162] Denise C Park. 2002. Aging, cognition, and culture: a neuroscientific perspective. *Neuroscience & Biobehavioral Reviews* 26, 7 (2002), 859–867.

[163] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 573–582.

[164] Nola Purdie and Gillian Boulton-Lewis. 2003. The learning needs of older adults. *Educational gerontology* 29, 2 (2003), 129–149.

[165] Kathryn Rindskopf and Don C Charles. 1974. Instructor age and the older learner. *The Gerontologist* 14, 6 (1974), 479–482. DOI:https://doi.org/10.1093/geront/14.6.479

[166] Sandra Sanchez-Gordon and Sergio Luján-Mora. 2013. Web accessibility of MOOCs for elderly students. In *2013 12th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 1–6. DOI:https://doi.org/10.1109/ITHET.2013.6671024

[167] Anne Shumway-Cook, Marcia A Ciol, Kathryn M Yorkston, Jeanne M Hoffman, and Leighton Chan. 2005. Mobility limitations in the Medicare population: prevalence and sociodemographic and clinical correlates. *Journal of the American Geriatrics Society* 53, 7 (2005), 1217–1221. DOI:https://doi.org/10.1111/j.1532-5415.2005.53372.x

[168] Thomas A Simonds and Barbara L Brock. 2014. Relationship between age, experience, and student preference for types of learning activities in online courses. *Journal of Educators Online* 11, 1 (2014), n1. DOI:https://doi.org/10.9743/JEO.2014.1.3

[169] Jaisie Sin, Rachel L. Franz, Cosmin Munteanu, and Barbara Barbosa Neves. 2021. Digital Design Marginalization: New Perspectives on Designing Inclusive Interfaces. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11. DOI:https://doi.org/10.1145/3411764.3445180

[170] Barbara A Soloman and Richard M Felder. 2005. Index of learning styles questionnaire. *NC State University* 70 (2005).

[171] Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. 2020. On Pause: How Online Instructional Videos are Used to Achieve Practical Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12. DOI:https://doi.org/10.1145/3313831.3376759

[172] Mojtaba Vaismoradi, Jacqueline Jones, Hannele Turunen, and Sherrill Snelgrove. 2016. Theme development in qualitative content analysis and thematic analysis. (2016).

[173] Pascal WM Van Gerven, Fred Paas, Jeroen JG Van Merriënboer, Maaike Hendriks, and Henk G Schmidt. 2003. The efficiency of multimedia learning into old age. *British journal of educational psychology* 73, 4 (2003), 489–505.

[174] Christina Victor. 2004. *The social context of ageing: A textbook of gerontology*. Routledge.

[175] Feliciano Villar and Montserrat Celdrán. 2013. Learning in later life: Participation in formal, non-formal and informal activities in a nationally representative Spanish sample. *European journal of ageing* 10, 2 (2013), 135–144.

[176] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 225–236. DOI:https://doi.org/10.1145/2858036.2858107

[177] Abeer Watted and Miri Barak. 2018. Motivating factors of MOOC completers: Comparing between university-affiliated students and general participants. *The Internet and Higher Education* 37 (2018), 11–20. DOI:https://doi.org/10.1016/j.iheduc.2017.12.001

[178] Maryanne Wolf. 2018. Skim reading is the new normal. The effect on society is profound. *Sat* 25 (2018), 09–41.

[179] Xiang Xiao and Jingtao Wang. 2017. Undertanding and detecting divided attention in mobile mooc learning. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2411–2415. DOI:https://doi.org/10.1145/3025453.3025552

[180] Jie Xiong and Meiyun Zuo. 2019. Older adults' learning motivations in massive open online courses. *Educational Gerontology* 45, 2 (2019), 82–93. DOI:https://doi.org/10.1080/03601277.2019.1581444

[181] Saelyne Yang, Jisu Yim, Aitolkyn Baigutanova, Seoyoung Kim, Minsuk Chang, and Juho Kim. 2022. SoftVideo: Improving the Learning Experience of Software Tutorial Videos with Collective Interaction Data. In *27th International Conference on Intelligent User Interfaces*. 646–660.

[182] Beste F Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 47–60.

[183] Tom Zhang, Michelle Taub, and Zhongzhou Chen. 2021. Measuring the Impact of COVID-19 Induced Campus Closure on Student Self-Regulated Learning in Physics Online Learning Modules. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 110–120.

[184] So-Yeon Ahn. 2019. Decoding "good language teacher"(GLT) identity of native-English speakers in South Korea. *Journal of Language, Identity & Education* 18, 5 (2019), 297–310.

[185] So-Yeon Ahn and Hyun-Sook Kang. 2017. South Korean university students' perceptions of different English varieties and their contribution to the learning of English as a foreign language. *Journal of Multilingual and Multicultural Development* 38, 8 (2017), 712–725.

[186] Minoo Alemi, Parisa Daftarifard, and Roya Pashmforoosh. 2011. The impact of language anxiety and language proficiency on WTC in EFL context. *Cross-Cultural Communication* 7, 3 (2011), 150–166.

[187] Mike Armstrong, Andy Brown, Michael Crabb, Chris J Hughes, Rhianne Jones, and James Sandford. 2016. Understanding the diverse needs of subtitle users in a rapidly evolving media landscape. *SMPTE Motion Imaging Journal* 125, 9 (2016), 33–41.

[188] Verena Bader and Stephan Kaiser. 2019. Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization* 26, 5 (2019), 655–672.

[189] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

[190] Rusty Barrett, Jennifer Cramer, and Kevin B McGowan. 2022. *English with an accent: Language, ideology, and discrimination in the United States.* Taylor & Francis.

[191] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[192] Louisa Buckingham. 2015. Recognising English accents in the community: Omani students' accent preferences and perceptions of nativeness. *Journal of Multilingual and Multicultural Development* 36, 2 (2015), 182–197.

[193] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[194] Changyan Chi, Qinying Liao, Yingxin Pan, Shiwan Zhao, Tara Matthews, Thomas Moran, Michelle X Zhou, David Millen, Ching-Yung Lin, and Ido Guy. 2011. Smarter social collaboration at IBM research. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 159–166.

[195] Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *arXiv preprint arXiv:2207.11345* (2022).

[196] Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Performance disparities between accents in automatic speech recognition. *arXiv preprint arXiv:2208.01157* (2022).

[197] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.

[198] Marko Dragojevic, Dana Mastro, Howard Giles, and Alexander Sink. 2016. Silencing nonstandard speakers: A content analysis of accent portrayals on American primetime television. *Language in Society* 45, 1 (2016), 59–85.

[199] Andy Echenique, Naomi Yamashita, Hideaki Kuzuoka, and Ari Hautasaari. 2014. Effects of video and text support on grounding in multilingual multiparty audio conferencing. In *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. 73–81.

[200] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User*

Interface Software and Technology (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. DOI:https://doi.org/10.1145/3472749.3474784

[201] Elizabeth Elliott and Amy-May Leach. 2022. False impressions? The effect of language proficiency on cues, perceptions, and lie detection. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* (2022).

[202] ETS. 2023. TOEFL iBT Listening Section. https://www.ets.org/toefl/test-takers/ibt/about/content/listening.html

[203] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122* (2021).

[204] Jairo N Fuertes, William H Gottdiener, Helena Martin, Tracey C Gilbert, and Howard Giles. 2012. A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology* 42, 1 (2012), 120–133.

[205] Ge Gao, Naomi Yamashita, Ari MJ Hautasaari, Andy Echenique, and Susan R Fussell. 2014. Effects of public vs. private automated transcripts on multiparty communication between native and non-native English speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 843–852.

[206] James D Harnsberger, Rahul Shrivastav, William S Brown Jr, Howard Rothman, and Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice* 22, 1 (2008), 58–69.

[207] Ari Hautasaari and Naomi Yamashita. 2014. Do automated transcripts help non-native speakers catch up on missed conversation in audio conferences?. In *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. 65–72.

[208] Helen Ai He, Naomi Yamashita, Ari Hautasaari, Xun Cao, and Elaine M Huang. 2017. Why did they do that? Exploring attribution mismatches between native and non-native speakers using videoconferencing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 297–309.

[209] Braj B. Kachru. 1992. World Englishes: approaches, issues and resources. *Language Teaching* 25, 1 (1992), 1–14. DOI:https://doi.org/10.1017/S0261444800006583

[210] Sei Jin Ko, Charles M Judd, and Diederik A Stapel. 2009. Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin* 35, 2 (2009), 198–211.

[211] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[212] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[213] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology* 46, 6 (2010), 1093–1096.

[214] Karen Livescu. 1999. *Analysis and modeling of non-native speech for automatic speech recognition.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[215] Roy C Major, Susan F Fitzmaurice, Ferenc Bunta, and Chandrika Balasubramanian. 2002. The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL quarterly* 36, 2 (2002), 173–190.

[216] Robert M McKenzie. 2008. The role of variety recognition in Japanese university students' attitudes towards English speech varieties. *Journal of Multilingual and Multicultural Development* 29, 2 (2008), 139–153.

[217] Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. Predicting human perceived accuracy of ASR systems. In *Twelfth Annual Conference of the International Speech Communication Association.*

[218] FMNSKL a AG NINAREH MEHRABI and Fred Morstatter. 2019. A Survey on Bias and Fairness in Machine Learning. *arXiv 1908.09635* (2019).

[219] Poppy Noor. 2021. 'I had to change who I am': 'bison' reporter Deion Broxton on his TV accent struggle. https://www.theguardian.com/us-news/2021/apr/02/deion-broxton-bison-montana-journalist-accent

[220] Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura. 2016. Estimating Speech Recognition Accuracy Based on Error Type Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 12 (2016), 2400–2413. DOI:https://doi.org/10.1109/TASLP.2016.2603599

[221] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[222] Mei-Hua Pan, Naomi Yamashita, and Hao-Chuan Wang. 2017. Task rebalancing: Improving multilingual communication with native speakers-generated highlights on automated transcripts. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* 310–321.

[223] Yingxin Pan, Danning Jiang, Michael Picheny, and Yong Qin. 2009. Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2353–2356.

[224] Yingxin Pan, Danning Jiang, Lin Yao, Michael Picheny, and Yong Qin. 2010. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1725–1734.

[225] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[226] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How Accurate Does It Feel?–Human Perception of Different Types of Classification Mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.

[227] ProPublica. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[228] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[229] Nobuhiro Shimogori, Tomoo Ikeda, and Sougo Tsuboi. 2010. Automatically generated captions: will they help non-native speakers communicate in english?. In *Proceedings of the 3rd international conference on Intercultural collaboration*. 79–86.

[230] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Interspeech 2019*. ISCA. DOI:https://doi.org/10.21437/interspeech.2019-1427

[231] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[232] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*. 934–938. DOI:https://doi.org/10.21437/Interspeech.2017-1746

[233] Kentaro Toyama. 2011. Technology as Amplifier in International Development. In *Proceedings of the 2011 IConference* (Seattle, Washington, USA) *(iConference '11)*. Association for Computing Machinery, New York, NY, USA, 75–82. DOI:https://doi.org/10.1145/1940761.1940772

[234] Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs.

[235] Peter Trudgill. 1997. *The social differentiation of English in Norwich*. Springer.

[236] Ron Van Buskirk and Mary LaLomia. 1995. The just noticeable difference of speech recognition accuracy. In *Conference companion on Human factors in computing systems*. 95.

[237] Irina-Elena Veliche and Pascale Fung. 2023. Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[238] Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry* 11, 8 (2019), 1018.

[239] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. DOI:https://doi.org/10.1145/1978942.1978963

[240] Yunhan Wu, Martin Porcheron, Philip Doyle, Justin Edwards, Daniel Rough, Orla Cooney, Anna Bleakley, Leigh Clark, and Benjamin Cowan. 2022. Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–12.

[241] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2020. See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In *22nd international conference on human-computer interaction with mobile devices and services*. 1–9.

[242] Lin Yao, Ying-xin Pan, and Dan-ning Jiang. 2011. Effects of automated transcription delay on non-native speakers' comprehension in real-time computer-mediated communication. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*. Springer, 207–214.

[243] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[244] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M Belsito, J Marc Overhage, and Clement J McDonald. 2004. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics* 73, 9-10 (2004), 719–730.

[245] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. 1–8. DOI:https://doi.org/10.1109/CIG.2018.8490433

[246] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

Curriculum Vitae in Korean

이름: 김서영

전자주소: youthskim@kaist.ac.kr

학력

2009. 3. – 2012. 2.	숙명여자고등학교
2012. 3. – 2016. 2.	연세대학교 컴퓨터과학과 (학사)
2016. 3. – 2018. 8.	한국과학기술원 전산학부 (석사)
2018. 9. – 2025. 2.	한국과학기술원 전산학부 (박사)

연구업적

1. Haechan Kim, Junho Myung, **Seoyoung Kim**, Sungpah Lee, Dongyeop Kang, and Juho Kim. “LearnerVoice: A Dataset of Non-Native English Learners’ Spontaneous Speech.” In Proc. Interspeech 2024, pp. 2325-2329. 2024.
2. Yoonsu Kim, Kihoon Son, **Seoyoung Kim**, and Juho Kim. “Beyond Prompts: Learning from Human Communication for Enhanced AI Intent Alignment.” CHI 2024 Workshop on Getting Back Together: HCI and Human Factors Joining Forces to Meet the AI Interaction Challenge, 2024.
3. **Seoyoung Kim**, Yeon Su Park · Dakyeom Ahn, Jin Myung Kwak, and Juho Kim. “Is the Same Performance Really the Same?: Understanding How Listeners Perceive ASR Results Differently According to the Speaker’s Accent.” Proceedings of the ACM on Human-Computer Interaction 8, no. CSCW1 (2024): 1-22.
4. **Seoyoung Kim**. “Investigating How to Design Inclusive Data-Driven Systems for Diverse User Groups.” In Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 153-155. 2024.
5. Yoonsu Kim, Jueon Lee, **Seoyoung Kim**, Jaehyuk Park, and Juho Kim. “Understanding users’ dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level.” In Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 385-404. 2024.
6. **Seoyoung Kim**, Donghoon Shin, Jeongyeon Kim, Soonwoo Kwon, and Juho Kim. “How Older Adults Use Online Videos for Learning.” In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2023.
7. Saelyne Yang, Jisu Yim, Aitolkyn Baigutanova, **Seoyoung Kim**, Minsuk Chang, and Juho Kim. “SoftVideo: Improving the Learning Experience of Software Tutorial Videos with Collective Interaction Data.” In Proceedings of the 27th International Conference on Intelligent User Interfaces, pp. 646-660. 2022.

8. **Seoyoung Kim**, Seokhun Jeong · Seulgi Choi, Juhoon Lee, Juho Kim. “Data-Driven and Personalized Vocabulary Recommendation for English Learners.” Korea Software Congress, 2021.
9. **Seoyoung Kim**, Arti Thakur, Juho Kim. Kim, Seoyoung, Arti Thakur, and Juho Kim. “Understanding Users’ Perception Towards Automated Personality Detection with Group-specific Behavioral Data.” In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-12. 2020.
10. **Seoyoung Kim**, Arti Thakur, and Juho Kim. “You are How You Behave in Your Group: Predicting Personality via Behaviors in a Co-located Group.” CSCW’19 Workshop on Learning from Team and Group Diversity, 2019
11. **Seoyoung Kim**, Sunwoo Kwon, Donghoon Shin, and Juho Kim. “An Analysis of K-MOOC Learners’ Data and an Investigation of its Future Applications.” Issue paper of Korea’s National Institute for Lifelong Education, 2019.
12. **S**eoyoung Kim, Jiyoun Ha, and Juho Kim. “Detecting Personality Unobtrusively from Users’ Online and Offline Workplace Behaviors.” Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018.
13. **S**eoyoung Kim, Taeho Kim, and Jinah Park, “Modeling 3D Cell Nucleus by Template-based Deformable Model with Confined-region Determined by Cellular Characteristics,” International Forum on Medical Imaging in Asia (IFMIA) 2017, pp. 17-20, 2017.
14. Jeongwoo Kim, Charndoh Bak, Inseop Kim, **S**eoyoung Kim, Junbum Cha, and Sanghyun Park, “SESE: Inferring disease-gene relationships using Second Sentence in biological literature”, Poster abstracts of The IEEE International Conference on Biomedical and Health Informatics (BHI 2016), LasVegas, USA, January, 2016.