EXPLORING THE IMPACT OF ROBOT- DELIVERED INTERPRETATION BIAS MODIFICATION ON DEPRESSED YOUNG ADULTS IN HONG KONG: A MIXED-METHODS ITERATIVE APPROACH

HUANG SHIMING

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

School of Design

Exploring the Impact of Robot-Delivered Interpretation Bias Modification on Depressed Young Adults in Hong Kong: A Mixed-Methods Iterative Approach

Huang Shiming

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

May 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Shiming Huang

ABSTRACT

Depression is a significant mental health issue affecting young adults in Hong Kong. Social robots offer a promising platform for delivering accessible and engaging interventions. This thesis explores the impact of robot modalities (text, audio, video) on user perceptions, experiences, and outcomes of an online imagery-enhanced elaborative interpretation bias modification (**eiIBM**) program for depressed young adults in Hong Kong. The research employs an iterative, scientific-centered approach across three studies. Study 1, a between-subjects experiment, examines differences in user perceptions, experiences, and outcomes between varied robot modalities and a no-robot control. Study 2, a within-subject interview, explores reasons for the differences or similarities found in Study 1 and elicits additional information on modality preferences. Study 3 improves the **eiIBM** program based on insights from Studies 1 and 2 and re-examines the effects of robot modalities. The findings contribute to theoretical models of technology acceptance for healthcare robotics (relational agents), inform design principles for effectively incorporating social robots into digital mental health interventions, and offer implications for developing accessible and research-centered robot-assisted therapies that promote cognitive resilience among vulnerable populations. This thesis advances knowledge on the nuances of how robot modalities shape user experiences and therapeutic outcomes, guiding the development of future AI-powered mental health solutions.

Keywords: social robots, depression, interpretation bias modification, young adults, Hong Kong

ACKNOWLEDGEMENTS

Throughout my academic journey at The Hong Kong Polytechnic University (PolyU), I have received an immense amount of assistance and support from various parties.

First, I would like to express my deepest gratitude to my chief supervisor, Prof. Hoorn Johan. I sincerely appreciate his thoughtful and insightful guidance throughout my entire Ph.D. journey. I am not an exceptional or brilliant student by Chinese standards, lacking a first-tier bachelor's background or a high academic achievement. Despite this, Prof. Hoorn Johan believed in me and provided unwavering support to help me become a researcher. Without him, I would not have had the opportunity to join the School of Design at PolyU and fulfil my dream. He not only influenced my research pursuits but also imparted his research and life philosophies, which will continue to nurture me for the rest of my life.

Additionally, I would like to express my profound gratitude to Prof. Lee, the Dean, and Prof. Siu, the (Research) Associated Dean of the School of Design. Under their exceptional leadership, our school has maintained its world-leading position. I also extend my thanks to the supporting staff in the School of Design, who have always been ready to assist me with various issues related to my research. I acknowledge the Laboratory for Artificial Intelligence in Design (AiDLab), Hong Kong Special Administrative Region, for providing me with an Innovation and Technology Fund to support my Ph.D. study.

My appreciation also goes to my peers and colleagues at PolyU. I would like to thank Ms. Yoyo Cheung Wing Yu, Mr. Stoney Wang Yiqiao, and Ms. Skye Wang Can for their support and assistance in my research. Ms. Wei Lai, Mr. Song Yao, Ms. Jen Yoohyun Lee, Mr. Sark Xing Pangrui, Ms. Caroline DU Yunhe, Ms. Zhang Yaqi, and Ms. Zhang Ya Ting also deserve thanks for their kind help in my research and life. I am not an outgoing person, but they took the initiative to approach me, making this Ph.D. journey a beautiful one.

Moreover, I am grateful to my sisters and brothers in the church, too many to list on a single page. Special thanks to Ms. Lai Ka Yi, Dr. Johnny Chan Yick Chun, Ms. Pun Yuen Ming, and Dr. Kristine Kit Yi Pang from the joyful fellowship, who have consistently encouraged and spiritually supported me throughout these years.

Furthermore, I appreciate the support from my family members: my mom, dad, little brother, elder sister, and my two nephews. Without their consideration and help, it would have been impossible for me to navigate this challenging yet fruitful journey. I also want to express my greatest appreciation to my helpful colleague, best friend, closest family member, and beloved husband – Dr. Johnny Ho Ka Wai. I understand that these three years have not been easy for him, but he has done his best to take care of me in every aspect of my life and research. He has cared for my every pursuit and devoted himself to helping me achieve them. I also thank the source of his love, God, who took the helm and turned my life around when I thought I would sink into the depths of the sea. Glory be to God.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 Background

1.1.1 Depression among Young Adults in Hong Kong

Depression is a major mental health problem that affects people around the world. According to the World Health Organization (WHO, 2019), depression affects more than 300 million people worldwide, accounting for approximately 3.8% of the population. The prevalence is slightly higher in women (6%) compared to men (4%). In Hong Kong, depression has become a major public health problem. Surveys indicate that 15.3% of Hong Kong residents report symptoms of anxiety disorders, while 12.1% report depressive symptoms (Ng et al., 2021). The young adults are the prominent sufferers. A study from Liu et al. (2022) presents data from an online survey conducted at Hong Kong University, revealing that 21% of students reported moderate to severe levels of depression and 41% of them anxiety. Another research (W.Y.Tam et al., 2023) indicates a worrying trend of increasing depression rates in adolescents, with 13% of females reporting moderate to severe depression. The findings suggest that adolescents in Hong Kong face higher prevalence rates of depression compared to their peers in Macau and mainland China, indicating a unique socio-culture context that may exacerbate mental health challenges.

Several factors contribute to Hong Kong's high depression rates, including work-related stress and sociopolitical changes (Tong et al., 2021; Hou & Hall, 2019). Over 30% of residents state that their job negatively impacts their mental health (Zhu et al., 2016). Passive events such as loss of employment, loss of family, or trauma also increase vulnerability to depression (Choi et al., 2020). Recent social movements, pandemic isolation, and economic decline have further escalated mental health issues, with 75% of residents experiencing some form of mental illness (Ni et al., 2020).

Depression is strongly associated with the risk of suicide, which is 25 times higher in depressed individuals compared to the general population (Lam et al., 2024). Each suicide can potentially trigger mental health problems in immediate contacts, propagating depression in communities, which is known as ripple effect (Cerel et al., 2016). However, many affected individuals do not receive adequate care. In 2022, only about 66,000 people were diagnosed with depression in mental health facilities in Hong Kong (Hospital Authority, 2023). Treatment costs and economic struggles also limit accessibility for this group already facing work challenges. The high prevalence of untreated depression inflicts damage on both individuals and society.

Research shows that mental disorders frequently first emerge in childhood, adolescence or early adulthood, with onset peaking around age 15 (Kessler et al., 2007). Depression is among the most common early-onset disorders. Globally, half of adults have experienced some mental disorder before 19 years old, and 75% manifest symptoms by 24 years of age (Kessler et al., 2005). This highlights the critical need for early intervention in young adults to mitigate the social costs of depression.

1.1.2 Target Population: Young Adults

As the evidence above shows, young adults are a high-risk group for emerging mental health issues such as depression. The developmental period termed ‘emerging adulthood’ spans ages 18 to 30 years (Arnett, 2024). During this time, people experience significant changes and challenges around identity, independence, relationships, and responsibility that can trigger mental health problems without proper support (Wood et al., 2017).

However, young adults also show low help seeking behaviors for mental health issues compared to other age groups (Aguirre Velasco et al., 2020). Barriers include lack of awareness about services, cost, stigma, and attitudes such as self-reliance. Other factors

include frequent transitional instability in emerging adulthood (Arnett, 2000; Stroud et al., 2015) and perceived inadequacy of available mental health services (Food and Health Bureau, 2017). The convergence of vulnerability, low help seeking, and inadequate care makes young adults an urgent target population for accessible and acceptable mental health interventions.

1.2 Social Robots

1.2.1 Definition

Social robots are distinguished from regular functional robots by their ability to interact and communicate naturally with humans in a social manner (Hoorn, 2015a). Although not possessing true sentience, they are designed to exhibit an 'agency' for socialization through capacities such as speech, empathy, and nonverbal behavior.

1.2.2 Types of Robots

Distinct types of social robots vary in their modalities and degrees of embodiment, which affects their capabilities and applications (Mollahosseini et al., 2018). Text-based (chatbot) or voice-based (audio bot) conversational agents simulate human dialogue using natural language processing and scripts. They lack physical embodiment. On the contrary, physically embodied robots can directly interact with their environment, but face mechanical constraints. Telepresence robots enable remote audiovisual communication through a platform containing a tablet screen, camera, speaker, and microphone. Physically embodied robots interacting with people through a screen could be regarded as a type of telepresence robot or on-screen robot. Virtual agents (avatars) are a type of voice-based agent with a virtual figure. They offer greater flexibility than telepresence and physically embodied robots but lack the tangible embodiments as chatbots and audio bots.

1.2.3 Robots in Mental Health

Robots offer advantages such as 24-hour availability, reduced stigma, customizability, and potential cost-effectiveness over time. Consequently, they are increasingly being explored as tools to increase the accessibility, acceptability, and affordability of mental health services (Khawaja & Bélisle-Pipon, 2023). The key advantage of social robots over traditional digital interventions, like applications or websites, is their capacity to engage users through interactive and empathetic interfaces (Guemghar et al., 2022; Karim et al., 2022). Social robots could encourage openness and honesty regarding mental health issues (Karim et al., 2022), which is important in therapy.

Most research has focused on socially assistive robots such as Paro (baby seal robot) as emotional companions for clinical populations. Studies show that they can improve mood, reduce anxiety or isolation, and simulate social interaction (Wada & Shibata, 2012; Guemghar et al., 2022; Karim et al., 2022). Although promising, these lack standardized therapeutic content.

Conversational agents, such as chatbots, are gaining traction for delivering structured psychosocial interventions through dialogue (Fitzpatrick et al., 2017; Inkster et al., 2018; Ly et al., 2017). For example, chatbots have provided cognitive behavioral therapy (CBT) and demonstrated effectiveness comparable to human therapists for some conditions (Fitzpatrick et al., 2017). However, chatbots in these studies lack the presence of robots in embodied form, nor the natural interaction as they used more rule-based communication strategies.

Some studies have evaluated chatbots to administer evidence-based therapy remotely (Li et al., 2023). However, research on robots with other modalities to deliver standardized therapies remains limited. Comparative studies evaluating the relative modalities merits and limitations of different modalities are needed.

1.3 Cognitive Behavior Therapy (CBT)

Cognitive behavioral therapy (CBT) is an empirically validated treatment that focuses on identifying and modifying dysfunctional thinking patterns that contribute to emotional disorders such as depression (Beck, 2005; Hollon, 2019). A key mechanism is targeting cognitive biases, automatic tendencies to process emotionally relevant information in a distorted manner (Mathews & MacLeod, 2005; Gotlib & Joormann, 2010).

One cognitive bias implicated in depression is interpretation bias, a tendency to interpret ambiguous information negatively (Everaert et al., 2012; Mathews & MacLeod, 2005). CBT aims to correct this bias through strategies such as reappraisal training. However, depressed individuals often struggle to generate positive alternative thoughts (Holmes & Mathews, 2009; Morina et al., 2011) and need help practicing frequently. This has motivated the development of more interpretation bias modification paradigms (IBMs) that train benign interpretations of ambiguous stimuli through repetition.

Research shows that online IBM programs delivered via text or audio can reduce negative interpretation biases and depressive symptoms (Karyotaki et al., 2017; Mira et al., 2017). However, high dropout rates remain a key limitation. Social robots represent a promising platform to make online IBM programs more engaging and effective for depression. However, research into robot-delivered IBM is lacking.

1.4 Lead in - Robot-Delivered Interpretation Bias Modification

This study explores the use of online social robots, such as chatbots, audio bots, telepresence robots, and avatars, as delivery platforms for an imagery-enhanced elaborative interpretation bias modification (**eiIBM**) paradigm designed to be easily accessible and low-cost for depressed individuals in remote settings.

Previous research indicates that elaborative IBM with imagery techniques can improve far-transfer effects on depression relief and therapeutic outcomes compared to automatic IBM approaches (Koster et al., 2017; Holmes et al., 2006, 2009). The structured nature of **eiIBM**, grounded in cognitive behavioral therapy principles, is well-suited for the delivery through conversational agents (Fitzpatrick et al., 2017; Fulmer et al., 2018).

Incorporating online robots into the **eiIBM** program aims to provide an interactive, intelligent platform that enables regular remote practice tailored for depressed young adults in Hong Kong. The portability and decentralized nature (i.e., therapy could be performed without human) of this approach can increase accessibility, engagement, and efficacy by allowing high-frequency sessions without limitations by time and location, with Artificial intelligence (AI)-enabled personalization.

Unlike traditional in-person therapy with self-guided homework, robot-assisted **eiIBM** provides ongoing support and interactivity during remote sessions to maintain motivation and tailor treatment. Evaluating the influence of various robot modalities on user experiences and outcomes can inform the optimal design of digitalized AI-augmented mental health interventions.

Furthermore, robot-assisted **eiIBM** aligns with the growing emphasis on scalable, preventive mental health interventions, particularly for at-risk youth populations. By leveraging robotics with artificial intelligence to deliver evidence-based cognitive training, this approach has the potential to provide an engaging and accessible way to build resilience and reduce the incidence of depression in vulnerable groups.

1.5 Research Objectives, Scope, and Questions

1.5.1 Research Objectives

The primary goal of this research is to investigate the impact of integrating social robots as relational agents on the user experience and outcomes of an online imagery-enhanced elaborative interpretation bias modification (eiIBM) program for young adults dealing with depression. The utilization of various modalities (such as text-based chatbots, voice-based chatbots, on-screen robots, and physically embodied robots) in mental health interventions has garnered attention due to their diverse effects on users. It is crucial to compare these modalities as they can produce different effects in terms of interaction capabilities and user engagement. Each modality offers unique interaction capabilities that enhance the user experience, ultimately influencing how users engage with the technology. For example, voice interaction may create a more personalized experience compared to text, affecting the therapeutic effectiveness, and social entities can foster trust and emotional connections. Our aim is to delve into the disparities in user experience within the realm of robot therapy. Therefore, the specific objectives are as follows:

1. To examine how user perceptions, experiences and intervention outcome differ between interacting with an **eiIBM** program delivered through varied robot modalities (audio, video, text) compared to a control condition without robots (Chapter 5).
2. To obtain additional information on subjective perceptions and preferences between robot modalities from a perspective within-subject (Chapter 6).
3. To improve the **eiIBM** program based on insights from Study 1 and 2 and re-examine the effects of varied robot modalities on experiences and intervention outcomes (Chapter 7).

1.5.2 Scope and Research Questions

The scope of this thesis is centered around investigating the impact of social robots on user experiences and outcomes when incorporated as delivery mediums into an imagery-enhanced elaborative interpretation bias modification (**eiIBM**) program for depressed young adults.

The specific research questions addressed are as follows:

For Study 1: How do different robot modalities (text, audio, video) influence user perceptions and experiences, and thus outcomes in an online **eiIBM** program? (**RQ1**)

For Study 2: What are the underlying reasons for the differences or similarities in user perceptions, experiences, and outcomes across robot modalities delivering **eiIBM**? And what additional insights can be gained regarding modality preferences? (**RQ2**)

For Study 3: What are the effects of the enhanced robot modalities on user experiences and outcomes in the refined variant of **eiIBM** that improved based on the findings from Studies 1 and 2?

The extension of the research questions for each study are described in Chapters 5, 6 and 7.

1.5.3 Significance of the Present Study

This study contributes to research on the incorporation of robots into psychotherapy and has potential theoretical, design, and social impact. From a theoretical perspective, the study extends the technology adoption models to the context of robot-assisted therapy by examining how user perceptions of different robot modalities influence engagement and outcomes of the **eiIBM** program (Chapters 5 and 7). By identifying the cognitive and affective determinants of robot use and engagement in the therapeutic setting, the findings contribute to refining theoretical models of technology acceptance for healthcare robotics

applications. In addition, the study advances knowledge on guiding principles for effectively incorporating social robots into digital mental health interventions, using the case of interpretation bias modification for depression.

In terms of design impact, the study compares user experiences and outcomes across different robot modalities (audio, video, text) in delivering an imagery-enhanced elaborative interpretation bias modification (**eiIBM**) program. This provides insight into how the choice of robot platform influences user engagement and therapeutic efficacy (Chapters 5 and 7). Qualitative exploration of user perceptions and preferences between robot modalities from a perspective within the subject (Chapter 6) offers a deeper understanding of the factors influencing robot acceptance in therapy, forming the design of research-centered robot-assistance interventions. Furthermore, improving the **eiIBM** program based on user insights and reexamining the effects of varied robot modalities (Chapter 7) demonstrates an iterative, research-centered approach to optimizing the design of robot-assisted therapies.

The current research could also inform broader guidelines for integrating AI and robotics into digital mental health solutions beyond just interpretation bias modification. By understanding the nuances of how different robot modalities shape user experiences and therapeutic outcomes, designers can make more informed choices about which robotic features to leverage for specific mental health applications. This study's insights into balancing robot features, engagement, and therapeutic efficacy can guide the development of future AI-powered mental health intervention that are both acceptable and effective for users.

From a social impact perspective, the study addresses the social challenge of countering negative bias in social cognition among vulnerable populations, particularly young adults highly exposed to social media (Khalaf et al., 2023) and at risk for depression. The **eiIBM** exercise could help build resilience in both depressed and non-depressed but vulnerable young adults by training them to interpret ambiguous online information (e.g.,

through social media) more positively, contributing to depression prevention efforts in society. Digitalization and decentralization of the **eiIBM** exercise enhance its portability and accessibility, allowing users to participate regardless of geographical location or time constraints. This expanded availability increases the potential frequency and consistency of usage, allowing the **eiIBM** exercise to reach and benefit a wider population. Decentralization also allows therapists to extend their services to multiple patients simultaneously, departing from traditional in-person therapeutic methods that require one-on-one sessions. This helps democratize access to therapy, overcome barriers in conventional practices, and promote mental health equity.

In summary, this thesis has theoretical, design, and social implications for the development of effective, accessible, and research-centered robot-assisted mental health interventions that address the social challenge of countering negativity bias and promoting resilience among vulnerable populations.

1.6 Thesis Outline

This thesis is organized into eight chapters as follow:

Chapter 1: Introduction – Provides background on depression among young adults in Hong Kong, social robots, cognitive behavior therapy, and the motivation for developing a robot-delivered interpretation bias modification program, followed by statements of research objectives, scope, questions, and significance.

Chapter 2: Literature Review - Reviews relevant theory and research on interpretation bias modification, social robots in mental health, and technology acceptance models. Finally, the research framework is proposed.

Chapter 3: Research Design and Methodology – Details the research paradigms, study designs, sampling methods, instruments, procedures, and analysis techniques used in the three studies.

Chapter 4: Test of Assessment Techniques – Describes the translation and psychometric validation of Chinese versions of the assessment tools used to measure variables of interest.

Chapter 5: Effect of **eiIBM_RobotV1**(Study 1) – A between-subjects experiment that examines differences in user perceptions, experiences and results between varied robot modalities (audio, video, text) and a waiting control in delivering the interpretation bias modification program.

Chapter 6: Experience of **eiIBM_RobotV1** Guided by Different Robots (Study 2) – A within-subject interview study exploring the reasons for differences or similarities found between robot conditions in Study 1 and eliciting additional information on modality preferences.

Chapter 7: Effect of **eiIBM_RobotV2** (Study 3) – Iterates on the program design based on insights from Studies 1 and 2 and re-examines the effects of the varied robot modalities on user perceptions, experiences, and outcomes using an improved version.

Chapter 8: General Discussion and Conclusion – Integrates and interprets the overall findings, discusses theoretical and practical implications, acknowledges limitations, and suggests directions for future research. Draws overall conclusions regarding the thesis objectives.

This outline provides a logical flow, with each study (Studies Chapters 5-7) building upon the previous one to address the research questions and objectives stated in Chapter 1. The literature review in Chapter 2 establishes the theoretical foundation, while the methodology in Chapter 3 ensures scientific rigor. The assessment validation in Chapter 4

supports the reliability and validity of the measures. The discussion in Chapter 8 synthesizes the findings into meaningful insights and implications.

Chapter 2: Literature Review

This chapter reviews previous research and theoretical concepts closely related to the present study. This review covers three significant aspects. The first section illuminates the therapeutic approach of the elaborative interpretation bias modification as the setting of my study by explaining the cognitive theory of depression, reviewing the cognitive processes underlying symptoms of depression, and comparing the interpretation bias modification paradigm. The second section reviews previous studies on robots in mental health. The final aspect is the theoretical models to explore the cognitive and affective determinants of use and engagement in robot-assisted therapy. The research framework was also proposed as the foundation for the thesis.

2.1 Elaborative Interpretation Bias Modification as a Therapeutic Approach to Depression

2.1.1 Cognitive Theory of Depression

Cognition has been central to conceptualizing depression for more than half a century (Beck, 1963; 1991). In the traditional Beck cognitive model (Beck, 1967), schema, namely the internally stored representation of experience, is considered to guide individuals to filter stimuli from internal or external experience so that information can be noticed and processed in a schema-consistent manner. Beck theorized that the schemas of depressed persons include negative themes such as loss, separation, failure, worthlessness, and rejection, which leads to specific bias in their attention to environmental stimuli and interpretation of information in a schema-congruent way (Beck, 1967, 1987, 2008; Clark et al., 2000). Kuiper et al. (1988) extended Beck's model by providing evidence that depressed people's self-schema consisting of negative content facilitated congruent information processing. More recently, Beck and Bredemeier (2016) proposed the unified model of depression, which acknowledged that

cognitive biases are the presupposition for the development and recurrence of depression. This model believes that genetic risk and stressful experience contribute to the negative self-schema and thereby initiate a vicious cycle of negative automatic thoughts, processing biases, and depressed mood.

Research examining processing biases in depression was inspired by Bower (1981), who first introduced the information-processing construct between mood and memory. In his associative network theory, memory comprises cognitive networks of numerous nodes, each containing a specific semantic representation that environmental stimuli can activate. The activation of any node causes partial priming of the other memory nodes within its associative network through diffusion. If memory nodes are frequently activated, the threshold to activate these associative networks will decrease. Therefore, frequent activation of associative networks with negative topics leads to negative information processing bias.

Based on Bower's (1981) work, Ingram (1984) further proposed that depression is the result of chronic activation of depression-mood nodes using the information processing paradigm. Individuals with such strongly activated depressed mood nodes have feedback loops where thoughts, memories, and associations related to depressed mood become more accessible for further processing. Because of the limited processing power, people with depression are impaired in their ability to get access to and process non-depressed related information, which deteriorates the depressed mood. Supporting information processing models (Bower, 1981; Ingram, 1984), a large amount of evidence shows that depressive adults have negative bias in attention, memory, and interpretation (Elliott et al., 2010; Gotlib & Krasnoperova, 1998). These studies also indicated that these processing biases are not only depression by-products, but rather stable factors of vulnerability to depression onset and relapse.

The association networks are distinguished from schemas by their inclusion of processing assumption (e.g., diffusion of associative nodes; see Anderson, 1976; Bower, 1981; Ingram, 1984) and exclusion of content assumption (i.e., they do not assume the structures to semantic memories/experience but assume that apparent structure is derived from how information is used) (Krawietz et al., 2012). However, they are considered interchangeable due to their functional similarity and their critical role in depression. In this paper, I will use the term schema.

Consistently, the dominant cognitive theories of depression postulate that the negative self-schema or the negatively active associative network is a stable construct, indicating that depressed individuals are expected to endure the attentional biases of internal and external stimuli beyond the depressive episode. Additionally, depicted as a component of the vicious cycle, biased information processing contributes to the onset, maintenance, and reoccurrence of depression.

2.1.2 Automatic and Elaborative Interpretation Biases

Emotional biases in attention, interpretation, and memory are viewed as critical cognitive processes that underlie the symptoms of depression (Everaert et al., 2014). The combined cognitive bias hypotheses revealed the interplay of these biases and the researchers found better effects on the intervention with all of them (Everaert & Koster, 2020). However, pathway analyses that tested the combined cognitive bias hypotheses found that negative interpretation bias was central to forming memory biases without the direct effect of attention biases on memory (Clark & Wells, 1995; Williams et al., 1988; Everaert & Koster, 2020). Many studies repeatedly demonstrated the centrality of interpretation biases in the formation of negative memories and, thus, depressive schemas (Everaert & Koster, 2020). Therefore, I

argue that compared to attention and memory biases, researchers should focus on interpretation biases for depression therapy purposes.

Rooted in Lazarus and Folkman (1984), Beck and Haigh (2014) distinguished automatic and reflective cognition systems in information processing. The automatic system processes stimuli rapidly (<1500ms), uses few cognitive resources, and is triggered by events that signal losses, threats, or gains. The threshold of response time is set according to the time limit in the automatic response priming paradigm (e.g., Affective priming: Tzavella et al., 2020; Inhabitation response: Ratcliff et al., 2018; Davranche et al., 2018; Stroop effect: Starreveld & La Heij, 2016). Reflective systems process information slowly (> 5000ms), are resource demanding and are controlled and deliberate. These systems may act reciprocally: The automatic system makes quick and subjective judgements, and the reflective system works on appraising, correcting, or modifying those judgements by re-evaluating them with time and resources (Beck & Haigh, 2014). Theoretically, such automatic biases instinctively prime schema-congruent responses. They can be corrected if the individuals become aware of their automatic thoughts and consciously devote cognitive resources to challenging the thoughts (Beevers, 2005). Although most of the studies that aim to modify cognitive biases without conscious awareness in depression demonstrated training-related improvements (near transfer), many of them lacked far transfer effects (i.e., less depressive symptoms or being less biased in real-life events) (Koster et al., 2017). Therefore, it could be possible that cognitive bias modification (CBM) procedures focused on improving elaborative rather than automatic processing modes could be more beneficial to disorders like depression, where these elaborative mechanisms seem to be more affected than automatic ones (Duque et al., 2015).

Compared to automatic interpretation processes, elaborative interpretation is more reflective. It reevaluates the initial conclusions drawn from automatic processes, correcting

the judgements as more information is integrated from either external attention (i.e., environment) or internal attention (i.e., memories of similar experiences) (Mathews, 2012; Ouimet et al., 2009; Wisco, 2009; Mathews & MacLeod, 2005). Meanwhile, elaborative interpretation biases highly intercorrelate with attention and memory biases (Joormann et al., 2015; Everaert et al., 2014; Everaert et al., 2013; Everaert et al., 2012). Given that cognitive theories have emphasized dysfunction at the automatic processing level in depression (i.e., automatic thoughts drive the disorder; Beck, 1979; Beck &Haigh, 2014; Beevers, 2005; Mathews & MacLeod, 2005), I argued that automatic interpretation bias characterizes the negative schema of depression, whereas elaborative biases could be the point of penetration to intervene the depressed thinking.

Although there were positive training effects of interpretation biases in the population, far-transferred effects on depressive symptoms were sometimes absent (e.g., Yiend et al., 2014). One probable reason is that the experienced effect covered the actual effect of the training tool. In most studies, the researchers used the same paradigm for training and assessment. Repeated training makes participants skilled in the pattern response, which might make the intervention tool falsely valid. Another probable reason is that the stimuli are different from the reality scenario. Therefore, participants cannot map cognition activity to environmental stimuli.

2.1.3 Automatic and Elaborative Interpretation Bias Modification

There are currently three main types of cognitive bias modification of interpretation bias (CBM-I): the semantic priming paradigm, the semantic association paradigm, and the ambiguous situations paradigm where the first two types target automatic biases, whereas the latter targets elaborative biases. All of them can be adapted to be assessment instruments by removing the feedback components.

2.1.3.1 Semantic Priming Paradigm

The homograph paradigm is one example of semantic priming paradigm-based CBM-I (Mathews et al., 1989). In this paradigm, participants are presented with words (most aurally) with the same pronunciation but different spellings and thus meanings, mostly one negative and one benign (e.g., die and dye). The participant will receive positive feedback if they recognize the word with benign meaning and vice versa. As in the iteration of the homograph paradigm CBM-I, Dearing and Gotlib (2008) created the Word Blends paradigm, in which participants listen to the ambiguous auditory stimuli that are constructed by acoustically blending two words that differed by only one phoneme (e.g., sad-sand). Participants are instructed to select the word they thought they heard from the two choices presented. They are given positive feedback if they choose the benign word and negative feedback if they choose the negative word.

CBM-Is based on the semantic priming paradigm train participants to confront the negative automatic response to priming stimuli by actively recalling the positive replacement. The general advantages of using word-level paradigms include their ease of administration (i.e., they can be presented without the use of a computer). However, a general limitation is the small number of appropriate word pairs available for the training sets. For example, there are relatively few homophones that have negative-related and neutral meanings, especially in the Chinese context.

2.1.3.2 Semantic Association Paradigm

The Word-Sentence Association Paradigm (WSAP) (Beard & Amir, 2009) was initially developed as an assessment instrument and has been adapted for the interpretation bias modification tool. As an example of semantic association paradigm-based intervention, it manipulates interpretation biases by providing positive feedback when participants associate

the benign word or disassociate the negative word with the ambiguous scenarios (text-based or image-based) and provides positive feedback in response to the negative interpretation of ambiguous scenarios. In this paradigm, a single word is presented immediately after the ambiguous stimulus is removed. In contrast to semantic priming paradigms (e.g., homophone paradigm), the semantic association paradigm (e.g., WSAP) presents an ambiguous stimulus without cues as to the possible meaning of the information. In a WSAP task, participants may first be presented with an ambiguous scenario, such as: "You are walking down the street and see a group of people laughing together." After the ambiguous scenario is removed, a single word is presented, such as "Rejected". Participants are then asked to indicate whether the word is related to the previous scenario by making a speeded decision. Individuals reveal a greater likelihood of the bias existence through the faster endorsement decision (i.e., indicating the sentence and the word are related) because a faster association reflects the fitness of the unambiguous word and the semantic model (i.e., semantic expectation) already formed by the individual.

Furthermore, higher endorsement rates that accept negative or reject positive interpretations reflect higher possibilities of negatively interpreting ambiguous materials. The researchers argued that since participants in WSAP tasks are generally required to respond within 1200 ms, the stimulus interpretation should be relatively automatic (Cowden et al., 2015; 2017). This intervention is designed to train participants in a more benign interpretation of the stimulus with repetition of trials. Removing the feedback component transforms it into the assessment instrument with two indices (i.e., endorsement rate and speed) used to determine the change in interpretation bias. Unlike the word stimulus, ambiguous association paradigms include much more detailed stimuli (i.e., explicit scenarios) that can be created and tailored to the targeted population. However, a potential limitation is that researchers need to know whether participants read each sentence presented and compare it to the single words or

whether they merely respond to the single words (Beard & Amir, 2009). Therefore, researchers reversed the stimulus order (e.g., Cowden Hindash and Amir, 2012; Cowden et al., 2017) and claimed that their subsequent resolution of ambiguity was less influenced by the judgement of the words. However, participants must still be guaranteed to read each sentence and participate in the task as intended. Including a comprehensive question could be a positive way to overcome the limitation.

2.1.3.3 Semantic Situations Paradigm

The ambiguous situation paradigm (e.g., Mathews & Mackintosh, 2000) is the most widely used protocol to modify interpretation bias. In the task, auditory/word descriptions of ambiguous scenarios are presented to the participants, followed by a final sentence with some missing letters of the positive word for the participants to fill out to resolve the ambiguity. Subsequently comes the yes or no comprehension question. Each trial ends with "correct" feedback if participants demonstrate comprehension of the benign interpretation of the scenario and "incorrect" feedback to the negative interpretation. In the standardized ambiguous situations paradigm, after all scenarios (e.g., 10) within the session have been viewed, there follows the recognition memory task, in which the participants are asked to rate how associated comprehension statements are with ambiguous scenarios. Figure 2.1 display an example of the task within the ambiguous situation paradigm. Those who rate the highest score for the statements correctly and benignly depict the corresponding scenario are rewarded with positive feedback. Several modified ambiguous situation paradigms have recently been created for better training effects. For example, instead of filling out the missing letters and selecting the comprehensive statements, Brettschneider and his colleagues (2015) asked participants to opt for the statement (from one positive, one neutral, and one negative interpretation) that most likely came to their mind and to rate the probability of each

interpretation and the potential cost if the first option in mind happened. This paradigm believes that conscious awareness and correction of the negative interpretation amplifies the modification effects. Brettschneider et al. (2015) also added an avatar to guide the participants to a positive interpretation and further encourage them to correct themselves by mimicking the avatar.

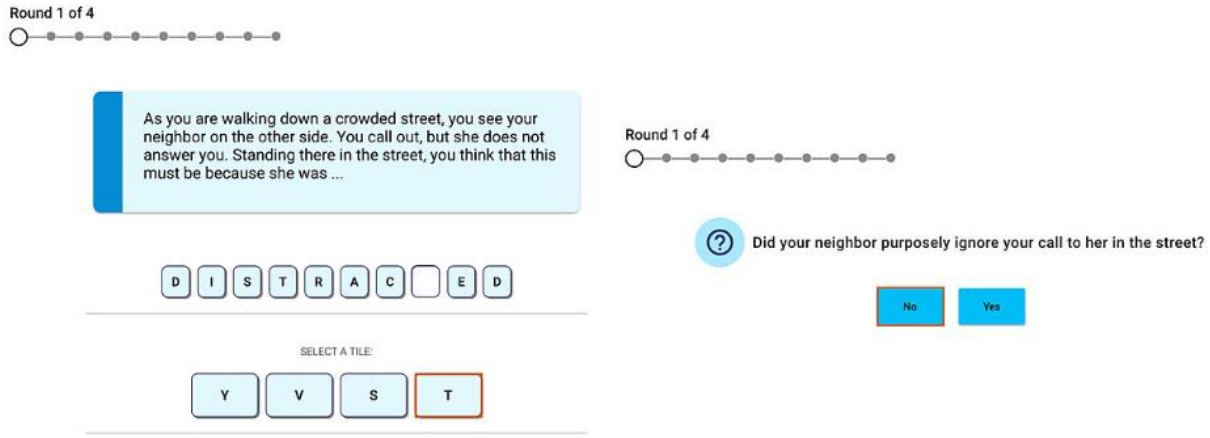

Figure 2. 1 An example of a CBM-I training scenario from MindTrails (Source: MindTrails)

The ambiguous situation paradigm aims to enforce participants' positive resolution towards ambiguous scenarios. Given that time allows for other information to be incorporated into the meaning assigned to the ambiguous stimuli, the ambiguous situation paradigm-based intervention modifies the elaborative interpretation biases which are associated with attention and memory biases (Everaert et al., 2012; Everaert et al., 2013; Joormann et al., 2015). Indeed, these paradigms considered the participant's memory of scenarios (e.g., the recognition part in Mathews & Mackintosh, 2000).

As an assessment and intervention instrument, there are controversies about the recognition accuracy as an index of interpretation biases and, thus, the training effects. First, it is hard to distinguish the elaborative interpretation biases from response styles (e.g., generally choosing the negative option) and expectancy biases (i.e., choosing options believed to be expected by the public). Second, it is challenging to disentangle elaborative

interpretation biases from other aspects of biased cognition. For example, elaborative interpretation biases could be falsely taken from attention and memory biases, given the evidence that depressed individuals tend to have a longer time to get away from negative environmental information compared to healthy controls (Kellough et al., 2008); for reviews (see Gotlib & Joormann, 2010; Teachman et al., 2012) and that they tend to recall negative information more easily (for reviews, see Gotlib & Joormann, 2010; Matthews & MacLeod, 2005).

However, I argue that it could be an appropriate intervention tool, as it mobilizes various cognitive processes that could actively change memory and benefit depressed adults. As self-perception theory posits, people determine their attitudes and preferences by interpreting the meaning of their behavior. The semantic situations-based paradigms include the components that encourage more self-reflection in their response, thus enforcing positive attitudes and preferences. For example, in the standardized ambiguous situation paradigm from Mathews and Mackintosh (2000), participants are asked to resolve the scenario ambiguity by filling in the missing fragment of the final word. With the internalization of attitudes and preferences, participants can transfer the effects to the new ambiguous stimuli. Therefore, it is supposed that the active assignment of meaning to ambiguous scenarios during training is critical for altering subsequent emotional responses to new stimuli (Hoppitt et al., 2010; Schweizer et al., 2011). However, there is a limitation in that participants need more autonomy to resolve ambiguous scenarios (i.e., the options are made by a computer).

In contrast, researchers found that the preference change was observed only when participants believed they had been instrumental in deciding (Lee & Daunizeau, 2019). Furthermore, these text-based paradigms have been reported to be relatively boring and labor intensive, requiring participants to read many lines of text (Beard et al., 2011). Improving

their autonomy and decreasing their reading workload could be attempted to make such paradigms more attractive and practical.

2.1.3.4 Imagery-Focused Paradigm

In addition to the three main types of modification that target the interpretation, a variant of modification with more imagery (e.g., Berna, Lang, Goodwin, & Holmes, 2011) has been developed and applied in the context of depression, with empirical evidence showing promising effects on interpretation biases and symptoms of depression (Bibi et al., 2020; Torkan et al., 2014; Rohrbacher et al., 2014; Lang et al., 2012). These imagery-focused paradigms make the participant repeatedly practice imagining positive resolutions for ambiguous situations, aiming to train them to automatically imagine positive resolutions for novel ambiguous situations encountered in daily life. Although the original scenario paradigms (i.e., WSAT and ambiguous situations paradigm) include an imagery component, they do not explicitly instruct participants to imagine the stimuli.

The mental images are internal representations "giving rise to the experience of seeing with the mind's eye" without an appropriate sensory input (Görgen, 2015). Mental representation of an object, activity, or experience can simulate experience and carry a similar emotional charge (Pearson & Kosslyn, 2013). Previous research has shown that processing stimuli using active imagery has stronger effects on interpretation bias and emotional vulnerability than processing the same stimuli verbally (i.e., reading the text; Holmes et al., 2009; Pearson et al., 2015), indicating that mental imagery could be an important target for treatment. Görgen (2015) explored whether mental imagery of positive and negative stimuli is superior to other processing modalities (e.g., verbal processing, pictures) in generating congruent affect in the explicit and implicit measure. It was found that depressed individuals might benefit from positive mental imagery with respect to the implicit or automatic level.

The effect of affect (i.e., emotion) influences the retrieval or gains of the resource for interpretations. Positive affect could suppress depressive schema activity and increase positive nodes' trigger ability in the association network. In other words, the promotion of positive imagery could be a treatment approach for depression.

However, depression has also been associated with difficulties in mental imagery of (future) positive events (Holmes et al., 2009; Morina et al., 2011). Moreover, depressed people are sensitive to negative stimuli, regardless of the processing modality (imagery, verbal processing, picture). There have been attempts to amplify the imagery effect in interpretation bias modification, such as adding explicit imagery instructions or doing the imagery description from a first-person perspective. These methods brought promising results in mood and interpretation biases. However, the limitation is that depressed people might report incongruent feelings during the training, as an intense sense of disagreement may emerge between the depressive schema (negative) and the imagery content (positive) (Kutas & Federmeier, 2000; Swaab et al., 2012).

2.1.4 Variants of Interpretation Bias Modification

In the previous subsections, I introduced the interpretation bias modification paradigm in depression that focuses on two cognitive targets: mental imagery and interpretation. Those targeted at the interpretation can be divided into automatic interpretation bias and elaborative interpretation bias. The paradigm focused on automatic processing is weak in transferring knowledge to the new stimuli. Thus, I would like to focus on the elaborative interpretation bias. Elaborative interpretation bias modification (i.e., ambiguous situation paradigm; **eIBM**) and the imagery-focused paradigm (**iIBM**) have been shown to achieve satisfying training results. While both **eIBM** and **iIBM** aim to change negative cognitive biases, they do so via different methods: **iIBM** via explicit 'top-down' cognitive evaluation and behavioral

experiments (Williams et al., 2015), and **eIBM** via a potentially more direct 'bottom-up' cognitive training approach (Williams et al., 2015). In the top-down approach, participants try to internalize the external information as their experience. In contrast, the bottom-up approach requires participants to handle the information from the sensory input repeatedly and update the initial system. However, **eIBM** has limitations in the considerable reading workload and less incongruent autonomy (they can only present as if they can interpret themselves). The pictorial-based approach and the self-generation component address the limitations mentioned and achieve better training results.

Pictures can give the participant more vivid scenarios than text, and self-generation encourages a stronger sense of substitution. In other words, these approaches improve self-relevance, which is essential in training, as it is the way to induce engagement that could facilitate the update of knowledge (Jin et al., 2019; Hess et al., 2009). In addition to that, researchers also noted that the familiar scenario could improve self-relevance (Wisco & Nolen-Hoeksema, 2010). Although the picture and self-generation components benefit **eIBM**, they do not facilitate **iIBM** (Rohrbacher et al., 2014). Those approaches damage the imagery process (c.f., Henricks et al., 2022; Rohrbacher et al., 2014). For example, self-generation gives people time to feel that everything is not real, and people with depressed moods may find it particularly difficult to generate positive interpretations (cf. Wisco and Nolen-Hoeksema, 2010). Furthermore, the unfamiliarity of images pulls participants away from engagement much more than verbal stimuli in mental imagery (Henricks et al., 2022).

Although limitations in themselves, these two modification approaches appear to be effective in adjusting interpretation bias and depressive symptoms. As depressed individuals struggle to imagine positive further events (Holmes et al., 2009; Morina et al., 2011), training positive interpretation (via consistently positive resolutions of the ambiguous situations; cf. Mathews & Mackintosh, 2000) in conjunction with positive mental imagery (via repeated

positive imagery generation) may therefore be particularly helpful in reducing the interpretation bias (both elaborative and automatic) and symptoms of depression, via targeting these particular processes synergistically (Holmes et al., 2009). This was not the first idea to integrate imagery-based modification with other treatments for depression. Blackwell and Holmes (2010) improved the imagery component of the **eIBM** by showing a relevant image before the scenario description. Williams et al. (2015) added positive imagery practice before The Sadness Program, a 10-week online course to counter sadness. They assumed that imagery practice equips participants to generate alternative thoughts quickly and facilitates anticipation of positive outcomes from behavioral tasks during the other treatment.

Compared to the condition with one of the putative active ingredients removed (i.e., no valence-specific training contingency was established by resolving half of the ambiguous training scenarios positively and the rest negatively), it was found that the imagery integrated into an existing treatment did not result in a significantly stronger reduction in the primary measures of depression and interpretation bias. However, a superior effect was observed in the follow-up measurement. Williams et al. (2015) discussed that the control condition contained many active components and may have presented a small dose of the positive condition. Despite this limitation, I argue that the positive imagery effect might be overrated. Depressed individuals struggle to disengage from negative stimuli. The control condition with an equal number of negative and positive imaginations could not have offset this effect.

The imagery likely made a limited contribution to the upcoming treatment. Depressed individuals lack the ability to imagine optimistic future scenarios. Another possibility is that interpretation with positive resolution (i.e., **eIBM**) may be an effective preparatory treatment module before engaging in **iIBM**. Accepting and generating positive solutions to ambiguous

scenarios could help prepare individuals for imagery involving experiences incongruent with their memory.

Evidence suggests that combining cognitive and behavioral treatment components is especially efficacious (Steinman & Teachman, 2014; McGillivray & Kershaw, 2013; Hofmann, 2004; Mattick et al., 1989). However, most of this research was conducted in the context of social anxiety.

In summary, in the thesis, I will integrate the imagery approach and the active elaborative interpretation to amplify the transfer effect on negative interpretation bias and depressive symptoms. I selected the Blackwell and Holmes (2010) paradigm, which integrated the ambiguous scenario paradigm (Mathews & Mackintosh, 2000) and the imagery paradigm (Holmes et al., 2006), which evidenced the potential as a standalone targeted interpretation for depression. The eIBM contains the component of active response generation, which is crucial for successfully modifying interpretation biases (Hoppitt et al., 2010; Mathews & Mackintosh, 2000). The **iIBM** further alters the schema through a mental image. To achieve this, I will improve the Blackwell and Holmes (2010) setting by showing the image throughout the scenario exercise and emphasizing the positive resolution relevant to the image at the end to increase their afterwards imagery effect. The variant of the interpretation bias modification in the thesis is named **eiIBM**.

2.1.5 Obstacles to Employing Integrated Paradigm

Simply combining **eIBM** and **iIBM** does not automatically address the challenges faced by each intervention when used separately.

First are the design deficits of these selected interventions. For example, the selected intervention paradigm (**eIBM**) lacks the autonomy for the participant to self-generate the resolution and contains a high reading workload. Also, the imagery (**iIBM**) requires the

participants to highly engage in activities which might be challenging to the depressed individuals.

Second, young adults seldom seek help for these trainings, which are primarily available in psychological environments, due to the social stigmatization of depression and the excessive cost of travel and finance. Internet-based intervention has been gaining traction from researchers to solve the issue. However, depressed adults were found lacking anhedonia during the online CBT experience (Blackwell et al., 2015).

The third challenge to ensure the integrated intervention effects is to maintain motivation for multiple training (i.e., avoid early dropout), yet the training with multiple sessions is regarded important for the intervention effect on the depressed individuals. The Feng et al. (2020) study trained the participants to generate vivid mental images and to spend a prolonged time imagining a positive ending. However, this imagery-enhanced paradigm did not positively affect the worried groups.

One possible explanation came from earlier work on the N400 brain potential, where the brain area reflects how easy it is to integrate information into a given context based on an individual's semantic memory (Kutas & Federmeier, 2000; Swaab et al., 2011). For that information discordant with semantic memory, it would be harder to integrate and is likely to violate one's expectations and thus produce a significant negative N400 amplitude. Therefore, a brief single session of interpretation training may not be sufficient to alter early-stage habitual interpretive processes.

Aligning with the spirit of the interpretation bias modification – repetitive exposure to the positive resolution to the ambiguous stimuli, it is believed that the exposure should be sustained in a period rather than a one-off. With the multi-session training, the patients have opportunities to apply the new interpretive style in real-world interactions, thus promoting the transfer of training to (depressive) symptom reduction. Also, the single-session CBM-I

studies are unable to address the sustained impact of changing interpretation bias and depression, as depression measures a period of state rather than a present one. Given multi-session online interventions, the typical challenges are to maintain the continuous user acceptance and Engagement of depressed individuals who are less motivated.

In summary, there were three obstacles to a stronger effect of the integrated paradigm on the depressed individuals - 1) the deficit of the **eIBM** and the **iIBM** for depressed individuals, 2) the high finance and travel cost to receive the multi-session training, and 3) the reduced motivation of depressed patients to do multiple-session training. As discussed above, the deficit in the therapeutic approaches for depressed individuals could be improved by enhancing imagery in **eIBM** (Blackwell & Holmes, 2010). The internet-based therapeutic platform has been promising to deal with the high-cost accessibility issue. With the popularity of the internet-based platform, robots are gaining interest in being integrated into the internet-based therapeutic platform to motivate the patients' use intention and engagement in the therapy.

2.2 Social Robots as Therapeutic Agents for Depression

This section introduces social robotics and its application in mental health intervention. Its potential to improve the **eiIBM** shall also be addressed.

2.2.1 Definition of Robots

The definition of social robots has been debated in human-robot interaction research. As Hoorn (2015a) discussed, social robots are distinguished from regular functional robots and software agents by their ability to interact and communicate naturally with humans in a social manner. Unlike basic AI aimed at utilitarian goals, social robots exhibit an "agency" for socialization (Hoorn, 2015a).

While social robots need not have free will or sentience like humans, they embody concerns and goals, allowing natural interaction similar to human partners (Hoorn, 2015a). It aligns with views of social robots as relational artifacts designed specifically for social-emotional roles beyond practical functions (Formosa, 2021).

In this thesis, social robots are considered collaborative interactional agents that assist developers in addressing issues like depression, rather than fully autonomous entities aiming to perform everything independently. Robots have a specific purpose to achieve, which is to collaboratively attain therapeutic goals through natural social interaction and relationship building with humans.

2.2.2 Types of Social Robots

Social robots are designed to interact with people in an interpersonal way, and they come in various forms with different modalities and degrees of embodiment. While some social robots have physical embodiments, others exist as virtual entities or software programs. These variations lead to differences in their capabilities and applications.

Chatbots, for example, are computer programs designed to simulate conversation through textual or voice interaction using natural language processing (NLP). They aim to provide interactions that like talking to a real person (Kevin et al., 2023). Voice-based chatbots are considered audio bots. Although chatbots do not have physical embodiments, they are considered a type of social robot in this thesis due to their ability to engage in human-like conversations (Edwards et al., 2016). Chatbots offer flexible conversational abilities unconstrained by physical limitations. However, their lack of embodiment reduces nonverbal cues.

Virtual avatars, while not robots in the strict sense, are digital entities commonly used in games and virtual realities (Mollahosseini et al., 2018). They can be considered as

components or substitutes of a robot, especially when embedded with natural language processing capabilities. Virtual avatars are entirely virtual and offer great flexibility in appearance and capabilities beyond physical constraints. However, their lack of physical embodiment can limit social presence and engagement (Bainbridge et al., 2010).

In contrast, physically embodied robots exist entirely in the physical world and can directly interact with their environment (Fosch-Villaronga & Drukarch, 2022). While their physicality enables tangible and multimodal engagement, it also imposes constraints on mechanics, appearance and mobility. For example, it requires intricate mechanical systems (Mollahosseini et al., 2018). Their physical existence may also raise user expectations for realistic emotional expressions and cause the expectancy violation effect (Go & Sundar, 2019).

Telepresence robots combine physical embodiment with virtual interaction by enabling communication through video chat. They blend physical existence with remote operation, distinguishing them from fully embodied robots and virtual avatars (Mollahosseini et al., 2018). Telepresence robots may achieve a balance of virtual flexibility and tangible physicality.

Telepresence and virtual robots offer comfort for technophobic users and allow suspension of disbelief for affective behavior. However, their lack of physical presence can reduce full engagement as compared to embodied robots. Still, with quality graphics and environment, virtual interactions can be highly engaging (Bainbridge et al., 2010).

Although the agencies mentioned above vary in physicality and visualization, which impacts their capabilities and modalities, they may be regarded as social robots that facilitate social interaction to some degree.

In this thesis, chatbots are specifically considered a type of social robot that are data-driven or AI-based, designed to engage in more natural conversational interactions compared

to rule-based chatbots with pre-defined responses. In contrast, telepresence robots in this thesis refer to on-screen representations of remote robots, allowing users to view and communicate with the robot via a computer or tablet screen, rather than physically embodied robots with human behind. In these cases, both chatbots and telepresence robots are the agent systems included in this thesis.

2.2.3 Prior Research on Robots for Emotions/Depression/Anxiety

2.2.3.1 Chatbots (Textual-Based and Voice-Based)

Chatbots and conversational agents are emerging as new tools for improving mental health and well-being. As artificial intelligence (AI) capabilities advance, researchers increasingly explore their potential as low-cost, accessible mental health interventions. Several studies have investigated chatbots' efficacy, especially for emotional problems like depression and anxiety.

Li et al. (2023) conducted a systematic review and meta-analysis examining conversational agents for mental health. Analysis of 22 studies on chatbots for depression, anxiety, psychological distress, and well-being found that chatbots outperformed control conditions in reducing symptoms. Chatbots employing cognitive-behavioral therapies were the most effective. The meta-analysis revealed small-to-moderate effects of chatbots on improving mental health outcomes.

One study illustrating chatbots' promise is chatbot Woebot, helped Japanese teenagers learn cognitive restructuring for managing anxiety (Nicol et al., 2022). Teens conversing with the empathic chatbot exhibited lower anxiety after learning to reframe worrisome thoughts positively. Similarly, Woebot delivering short daily CBT lessons decreased anxiety and depression in college students over two weeks better than an information-only control (Fitzpatrick et al., 2017).

Other research found that chatbots delivered psychotherapy effectively. For instance, Wysa mimicked an empathic therapist, demonstrating active listening and reflection (Inkster et al., 2018). In a trial, clinically depressed individuals interacting with Wysa experienced reductions in symptoms like low mood, anhedonia, and suicidal thoughts (Inkster et al., 2018). It exemplifies conversational agents' capacity to build therapeutic alliances.

The chatbot Shim provided a brief intervention based on positive psychology principles and cognitive-behavioral techniques, resulting in lower depressive symptoms in young adults after two weeks compared to waitlist controls (Ly et al., 2017).

Besides the text-based chatbots, voice-based and multimodal (text + voice) bots were beneficial in emotional relief. XiaoE, a CBT-based chatbot developed for depression and interacting with users through text, image and voice, compared to the e-book and Xiaoai, a chatbot designed to cater to small talk among general users, was found to significantly reduce depressive symptoms (He et al., 2022). A study by Ogawa et al. (2022) found that an artificial voice-based chatbot may positively affect the smile and speech in patients with Parkinson's disease.

These studies underscore that text- and voice-based chatbots can be promising technology in mental health. In particular, chatbots can deliver emotion regulation strategies (e.g., CBT) through conversational micro-intervention, with outcomes surpassing information provision alone.

The research on the potential use of AI chatbots in mental therapy has shown promising results, but it is important to acknowledge the limitations. Schick et al. (2022) argue that current studies lack proper randomization and control groups, weakening their findings. Molli (2022) and Bendig et al. (2022) echo similar concerns, emphasizing the need for larger sample sizes and more rigorous experimental designs.

2.2.3.2 Telepresence Robots and Physically Embodied Robots

Studies into the potential health benefits of social robots have shown some promising results, but only in a limited number of settings (Robinson et al., 2019), far from the numerous studies exploring chatbots' potential in mental health. According to Li et al. (2023), only five research papers since 2018 have used telepresence or physical robots as the delivery platform to interact with people with negative moods. Moreover, none of them delivered evidence-based therapeutic approaches.

Early work by Wada and Shibata (2007) found that the therapeutic robot PARO, designed as a baby harp seal, improved mood in older adults with cognitive deficits. PARO was shown to reduce depressive symptoms and loneliness among diverse elderly population in nursing homes (Jøranson et al., 2016; Robinson et al., 2013). Also, other studies found that PARO eased anxiety, isolation, and negativity in clinical populations like hospitalized children (Sabanovic et al., 2013), dementia patients (Moyle et al., 2018), and psoriasis patients (Law et al., 2022), while increasing positive mood through simulated social interaction. Besides PARO, other zoomorphic robots like the dog AIBO and the dinosaur Pleo facilitated positive emotions and lowered negative feelings in the elderly (Jung et al., 2017). These studies demonstrate that animal-like robots may provide a certain degree of emotional assistance.

Humanoid robots may also effectively regulate moods. The IVEY robot mitigated anxiety and depression in pediatric cancer patients by providing positive social support during treatment (Trost et al., 2020). Pepper improved high-school students' positive affect and self-esteem, reducing anxiety levels by interacting naturally (Amani et al., 2014). The social robot named EMYS improved workers' mental health by engaging in social

interactions and promoting stress-reducing activities, supporting the potential of using robots to enhance employee well-being in the workplace (Lopes et al., 2023).

Nao also delivered cognitive reappraisal training for children to reinterpret stressful events positively. Children interacting with Nao reported lower test anxiety than cartoons or written instructions (David & David, 2022). Also, Luo et al. (2022) found that university students benefitted more from robots compared to social media disclosure and dairy writing in negative emotion recovery. In other words, physical robots are suitable intervention platforms.

Other research examined factors enhancing robots' emotional impact, like personalization and physical presence. For instance, embodied telepresence robots offering personalized greetings were preferred by the students, eliciting more disclosure versus video agents (Powers et al., 2007).

These efforts to explore the incorporation of physical robots into mental healthcare reveal physical robots' promising potential for delivering emotional support and relief. However, incorporating robots into mental healthcare has yet to involve evidence-based interventions. Most of the existing research has focused on the effects of robot companionship, where the robot plays the role of an emotional companion but does not deliver a structured therapeutic.

Also, these efforts explore the effect of a physically embodied robot. Telepresence robots remain largely unexplored. However, telepresence robots offer both cost-effectiveness and scalability, extending the accessibility of therapeutic services, and they have exciting potential to contribute to online mental health therapy.

In summary, a research gap remains in investigating social robots, especially telepresence robots, as platforms for delivering standardized, empirically validated therapies to improve mental health. Well-designed studies incorporating robots into established

interventions like CBT are needed. Comparing the delivery of the eiIBM intervention across three main modalities - text, audio, and visual - can clarify their relative strengths and suitability. Specifically, text-based chatbots represent the text modality, audio bots represent the audio modality, and telepresence bots/avatars represent the visual modality. Evaluating the efficacy of the **eiIBM** program when delivered through these diverse modalities will help elucidate the unique advantages and limitations of each approach. It will explain how social robots can enhance emotional well-being through therapeutic processes rather than isolated companionship effects.

2.2.4 Social Robots and AI Integrated Interventions

To address the motivation issue in the integrated paradigm **eiIBM** (**eIBM** + **iIBM**), social robot incorporating natural language processing (NLP) capabilities could be beneficial. NLP enables social robots to engage in more human-like conversations, potentially increasing user motivation and engagement. Social robots can provide low-intensity behavioral interventions, such as one-on-one therapy, which is particularly valuable in many countries facing a shortage of healthcare staff (Moerman et al., 2018; Robinson et al., 2019). There are two attributions to the robot-delivered **eiIBM** strengthened by NLP.

Firstly, the robots' artificial intelligence improves the participant's autonomy in the **eIBM**. The Traditional interpretation paradigm requires the users to select the most aggregable option from the computer-generated choices, which might weaken the autonomy and cause expectancy biases (i.e., choosing options believed to be expected by others) with choices. It is easy to implement but may need to be a better approach. Researchers observed only when participants believed they had been instrumental in deciding that the preference change happened (Lee & Daunizeau, 2019). Allowing the users to generate their interpretation and guiding them to positive interpretation would improve the effectiveness of

eIBM. Robots with communication functions make this autonomy feasible. In **eIBM**, instead of selecting the fragment of the final word and rating the possibility of the resolution, incorporation of a robot enables the participants to input the positive resolution themselves, and the robot gives feedback by assessing the valence and the de-ambiguity of the answer.

Furthermore, social rewards from robots can motivate the self-generation of positive solutions through appropriate strategies - in other words, motivate engagement. Social interaction is essential to mental health, and one of the significant motivators for social interaction in human beings is the desire to gain social rewards (Kawamichi et al., 2016). Social reward refers to social approval, belonging, and social support. The appraisal during intervention could be regarded as social rewards, and the encouragement from social agents could stimulate the pursuit of social rewards. Depression is strongly associated with abnormal processing of social rewards at both behavioral and neurological levels (Bishop & Gagne, 2018). On the neurological level, the researchers found that depression severity and the level of activation of the human reward system, including dopaminergic neural circuits, are negatively correlated (Russo & Nestler, 2013). This results in behavioral abnormalities that are blunted responses to social rewards (i.e., social anhedonia) (Kupferberg et al., 2016; Forbes & Dahl, 2011; Silk et al., 2021).

Consequently, depression inhibits an individual's motivation to engage in social interaction as they fail to expect social rewards from humans. It supports the previous finding that depressed individuals struggle to imagine an optimistic future scenario (Holmes et al., 2009; Morina et al., 2011). In this case, humans are not the appropriate agent for delivering positive social feedback. In recent studies on non-clinical populations, the social reward processes of people with depression show patterns in person-to-person different from human-robot interactions (Zhang et al., 2021). Social rewards during human-robot interactions were less affected by depression than those during human-human interactions. Remarkably, people

with depression are more likely to expect and receive social rewards from robots than from humans under some circumstances. Consistent with these findings, social robots scored higher on persuasiveness (Lopez et al., 2017; Ghazali et al., 2019). In traditional **eIBM**, only "correct" and "incorrect" text feedback is given, and versions offered to depressed groups did not consider the failure of social rewards in these groups. Also, when the program disagrees with certain responses, depressed individuals show higher sensitivity for negative feedback when processing an external stimulus (Mueller et al., 2015).

Incorporating social robots into online eiIBM interventions may present a promising approach to address the unique challenges faced by young adults with depression in seeking and engaging with mental health support (Aguirre Velasco et al., 2020). Compared to traditional text-based programs or human-delivered therapy, social robots can offer a more accessible, relatable, and less stigmatizing mode of support for this target population. By harnessing the benefits of artificial intelligence and social rewards (Robinson et al., 2019), social robots have the potential to directly mitigate the common motivational and autonomy-related issues observed in eiIBM programs. The conversational and interactive nature of social robots can help reduce barriers to engagement and increase motivation, while their ability to provide non-judgmental social rewards and avoid directly denying users' responses can mitigate negative emotional reactions, making them a particularly promising approach for supporting this vulnerable population.

Inspired by these findings, I propose to integrate social robots into the online eiIBM program. The robots' capabilities can be leveraged to guide and encourage depressed young adult participants in the self-generated positive interpretation exercises, thereby enhancing the effectiveness of the eiIBM intervention for this target group.

2.3 Models of Technology Use and User Experience

If the goal of a robot-delivered (online) program is to be used as part of a therapeutic intervention, integration with individual therapy is essential to ensure the effectiveness of the intervention itself. However, merely incorporating evidence-based practices into the program does not automatically make them evidence-based. The essential aspect is addressing engagement during the design and development of the program incorporated in the robot and, further, the program's adoption or intentions to use, as all efforts may be in vain if no one substantially uses the app.

To understand the participants' experience of an artificial agent delivering the evidenced-based practices (**eiIBM** as a therapeutic setting in my study) and, therefore, understand the effect of this experience on therapy effectiveness, the Interactively Perceiving and Experiencing Fictional Characters (I-PEFiC) framework lies at the basis of my approach (Van Vugt et al., 2009). The Technology Acceptance Model series (TAMs; Davis,1986; Venkatesh & Davis, 2000; Venkatesh & Bala, 2008) and Unified Theory of Acceptance and Use of Technology series (UTAUTs; Venkatesh et al., 2003; Venkatesh et al., 2012) are models that explain individual technology adoption by determining the technology's factors and reasoning. However, unlike pragmatic tools, social robots exhibit human-like relational qualities that can elicit emotional responses from users, not necessarily related to technical functionality and reasoning. Therefore, the I-PEFiC framework was developed to explain adoption of relational artifacts like robots through both cognitive and affective determinants. Next, I introduce the technology adoption models that are more conventional after which I explain how I-PEFiC borrows from and elaborates on these older approaches.

2.3.1 Model Explanation

2.3.1.1 Technology Acceptance Model Series (TAMs)

The Technology Acceptance Model (TAM), as proposed by Davis (1989), rests on the theory of reasoned action (TRA) proposed by Fishbein and Ajzen (1980). TAM posits that the *Intention to Use Technology* (*IUT*) is primarily influenced by *Perceived Ease-of-use* (*PEOU*) and *Perceived Usefulness* (*PU*). However, TAM had its limitations in terms of explanatory power (*R2*) (Bayraktaroglu et al., 2019). It led Venkatesh and Davis (2000) to expand TAM into TAM2. TAM2 aimed to include additional determinants of *PU* and *IUT*, considering the user's experience with the system over time. This expansion aspired to retain the original TAM constructs while enhancing its ability to predict user behavior. Despite advancements, there was still room for improvement, which led to the development of TAMS by Venkatesh and Bala (2008). This model addressed *PEOU* constructs that are absent in TAM2. TAM3 presented a comprehensive network of determinants that influence IT system adoption and included anchor and adjustment determinants to reflect the user's experience over time. Figure 2. 2 showed the relationship among TAM 1, 2 & 3.

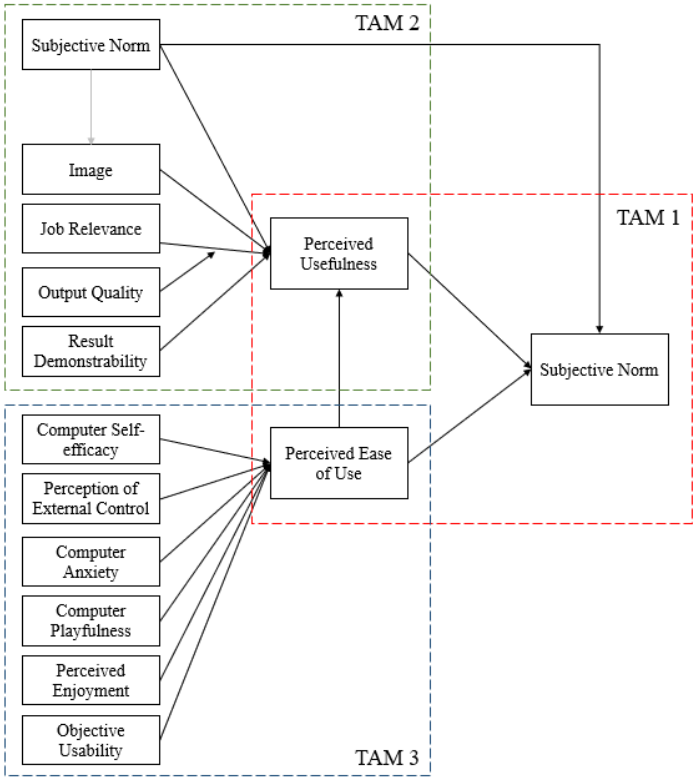

Figure 2. 2 TAM 1, 2 & 3 – Simplified omitting moderators, Davis (1989), Venkatesh and Davis (2000), Venkatesh & Bala (2008) (derived from: https://acceptancelab.com/technology-acceptance-model-tam)

2.3.1.2 Unified Theory of Acceptance and Use of Technology (UTAUT)

Meanwhile, a different approach was taken by Venkatesh et al. (2003) with the introduction of the Unified Theory of Acceptance and Use of Technology (UTAUT) model. UTAUT incorporated *Perceived Usefulness*, *Performance Expectancy*, and *Social Influence* as key determinants of *Behavioral Intention*, which together with *Facilitation Conditions* predict the actual use of the technology. Performance Expectancy is similar to TAM's *Perceived Usefulness*, and *Effort Expectancy* is similar to TAM's *Perceived Ease-of-use*. *Social Influence* and *Facilitating Conditions* are additional factors that are not explicitly included in TAM. Additionally, it also included four moderating factors: gender, age, voluntariness, and experience. The emphases of UTAUT are predicting technology adoption in social or organizational contexts and considering the demographics difference. However,

as Bagozzi (2007) pointed out, UTAUT emphasizes the mediating role of behavioral intention, potentially neglecting direct relationships that might exist outside the model's core constructs.

To address the limitations of UTAUT, Venkatech et al. (2012) further developed UTAUT2. This model was specifically designed to predict technology acceptance in consumer use, for instance in García de Blanes Sebastián et al., 2022; Yuan et al, 2015; Salgado et al, 2020. It added three additional constructs to the original UTAUT model: *Hedonic Motivation*, *Price Value*, and *Habit*. Finally, it consists of 7 construct variables including *Performance Expectancy*, *Effort Expectancy*, *Social Influence*, *Facilitating Conditions*, *Hedonic Motivation*, *Habit*, and *Price value*. These constructs were moderated by *Age*, *Gender,* and *Experiences*, extending the application of the UTAUT2 model (see Figure 2.3).

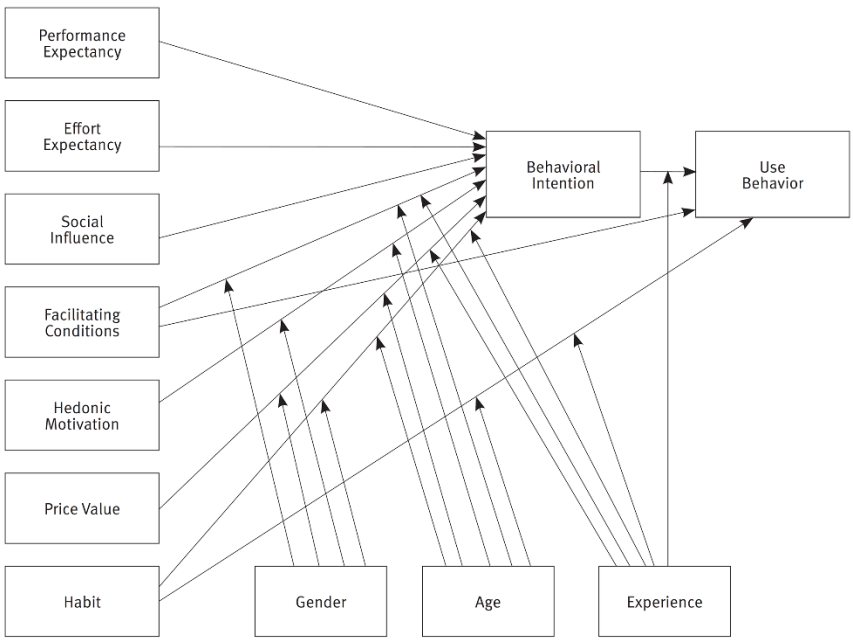

Figure 2. 3 Unified Theory of Acceptance and Use of Technology 2 (UTAUT2, Venkatech et al., 2012)

2.3.1.3 Extending Technology Adoption Models with TTF and EDM

Researchers have recognized the lack of affective factors in technology adoption models like TAMs and UTAUT2. Thus, Task-technology fit (TTF; Goodhue and Thompson, 1995) model and Expectation Disconfirmation Model (EDM; Oliver & DeSarbo 1988; Yi 1990) are often integrated to explain adoption intention and engagement more comprehensively.

2.3.1.3.1 Task-Technology Fit (TTF)

The TTF model (Figure 2.4) elegantly encapsulates the importance of task-contingency for evaluating technology performance and technology use intention (Goodhue & Thompson, 1995). It comprises of three key concepts: 1) **Task** - the activities users need to complete; 2) **Technology** - tools or systems employed during task execution; 3) **Fit** - the relevance and consistency between the task and the technology.

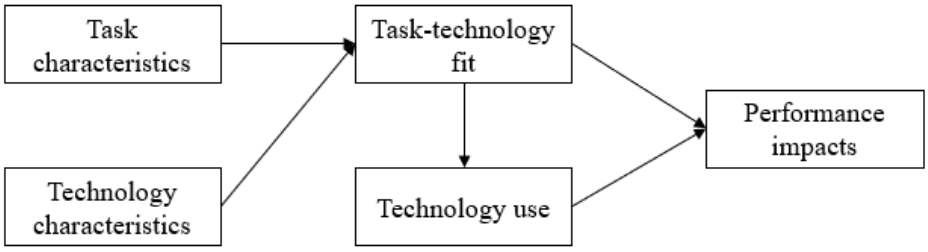

Figure 2. 4 Task-technology Fit (TTF; Goodhue and Thompson, 1995)

TTF posits that when a technology effectively supports the needs and objectives of a task, a good fit between the two is achieved. This fit leads to greater user satisfaction and improved task performance (Goodhue and Thompson, 1995; Goodhue, 1998).

For instance, in customer service robots, conveying effective information is a more fitting function than providing superfluous expressiveness (cf. Hoorn & Huang, 2024). This example highlights that TTF emphasizes evaluating fit with user needs beyond just technological functions. TTF has been integrated into prevalent technology adoption models

like TAM and UTAUT to explain how a technology's capability in fulfilling user's required tasks enhances performance expectancy and use intentions.

Perceived usefulness in TAM measures the expected performance impact of technology use, similar to performance expectancy in UTAUT. Studies confirm that TTF positively influences perceived usefulness, supporting their correlation (Dishaw & Strong, 1999). Therefore, TTF was integrated in TAM as the predictor of *Perceive Usefulness* (*PU*) and *Perceive Ease-of-use* (*PEOU*), addressing limitations that users may accept technologies despite unfavorable attitudes if it increases task performance (Letchumanan & Tarmizi, 2011; Goodhue & Thompson, 1995).

Similarly, when incorporated into UTAUTs, TTF positively influences *Performance Expectancy*, *Adoption Intention* (same as *Behavioral Intention* in UTAUT2), and *Use Behavior* (in UTAUT2) (Zhou, Lu & Wang, 2010; Tam et al., 2018; Paulo et al., 2018).

2.3.1.3.2 Expectation Disconfirmation Model (EDM)

The EDM (Figure 2.5) posits that expectations serve as a benchmark for evaluating performance and determining satisfaction/dissatisfaction (Erevelles & Leavitt, 1992; Oliver, 1997). EDM holds that satisfaction depends on the gap between expected and actual experience with a product, service, or technology.

Expectations refer to assumptions or predictions about qualities and features. **Perceived performance** means appraising actual performance. **Disconfirmations** denotes comparative judgement of perceived performance versus initial expectations. **Satisfaction** presents the contentment after gaining first-hand experience. Exceeding expectations brings positive disconfirmation and often increased satisfaction. Not meeting expectations causes negative disconfirmation and typically decreased satisfaction.

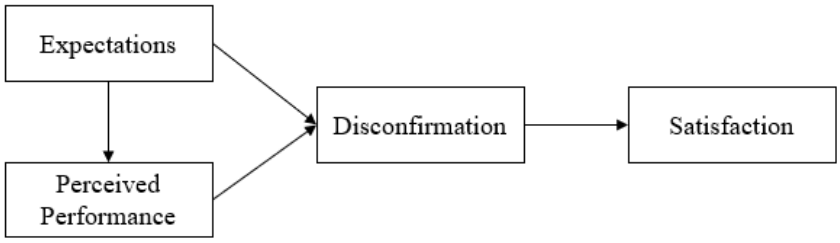

Figure 2. 5 Expectation Disconfirmation Model (EDM, Oliver, 1997)

EDM complements technology adoption models regarding the role of expectations. For instance, UTAUT2 includes expectation expectations like Performance Expectancy but does not directly address disconfirmation effects on affective side like Satisfaction.

Integrated into UTAUT2, EDM demonstrates that positive disconfirmation enhances performance expectancy and satisfaction, while satisfaction predicts actual use behavior (i.e., engagement, Tam et al., 2018; Singh, 2020). Though Satisfaction is not originally included as a construct in UTAUTs, it has been incorporated in many extended UTAUTs studies to provide richer affective and experiential considerations (Tamilmani et al., 2020; Chao, 2019). For example, in the extended UTAUT2 model, Satisfaction has been positioned as a key antecedent influencing users' continuance intention and actual use behavior (Tam et al., 2018; Singh, 2020).

2.3.1.3.2 UTAUT 2 Incorporated with TTF and EDM

Taken UTAUT2 as the representant of the technology adoption models, the integration of TTF and EDM into this model provides a more comprehensive explanation of technology adoption by considering the role of affective factors. The model suggests that when a technology is a good fit for the task that a user is trying to accomplish and when the user's expectations about the technology are met or exceeded, the user will experience positive affective outcomes, which influence the user's intention to use the technology and their actual use of the technology. To better display how TTF and EDM complement the

UTAUT2 model, I integrated the models through the shared variables among the models (Figure 2.6).

Figure 2.6 depicts the incorporation based on the literature review, where UTAUT2 factors are represented by the blue box on the right side of the model. EDM factors are represented by the green box on the left-top side of the model and TTF factors on the left-bottom side. The factors in red color are unapplicable in the thesis emphasizing private interaction. The arrows in the model show the relationships between the different model constructs.

The TTF consists of the predictors of *Performance Expectancy* (*PE*) and *Effort Expectancy* (*EE*), where *PE* is similar to the concept of Performance Impact in the original TTF model. *Performance Expectancy* (*PE*) and *Effort Expectancy* (*EE*) lead to *Behavioral intention*. *Disconfirmation* in a predictor of satisfaction, which is the feeling of contentment with a technology. *Satisfaction* together with the UTAUT2's original constructs predict *Behavior Intention* to use the technology.

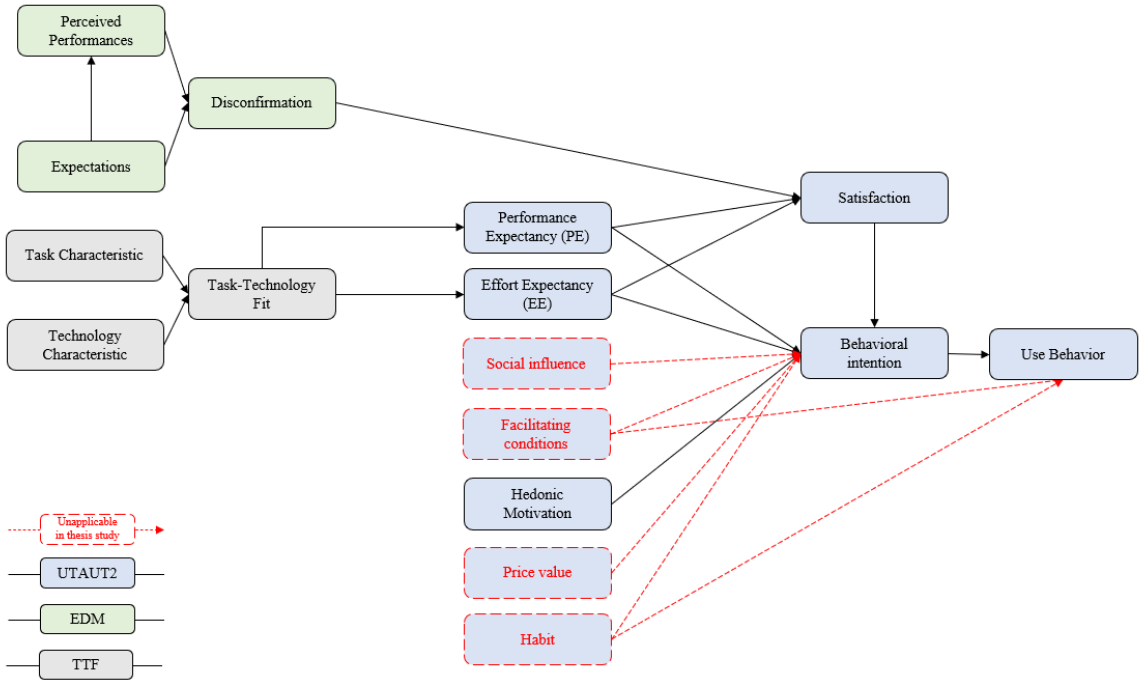

Figure 2. 6 Integration of EDM and TTF into UTAUT2 – Omitting Moderators.

2.3.1.4 Interactively Perceiving and Experiencing Fictional Characters (I-PEFiC)

Unlike TAM series and UTAUT series, which emphasize cognitive and rational perspectives in explaining user adoption of general technologies, I-PEFiC was developed to explain user responses to relational technologies like agent systems and social robots (e.g., Hoorn, Konijn, & Pontier, 2018). I-PEFiC incorporates stronger emotional and experiential aspects compared to mere-pragmatic adoption models.

In Figure 2.7, I-PEFiC unfolds across three stages when interacting with virtual agents like social robots. First, in the Encoding phase, individuals perceive attributes of the virtual entity, including its action potential (*affordance*), ethical qualities (*ethics*, good-bad character), *aesthetics* (attractive - unattractive), and human-likeness (*realism*). Next, in the Comparing phase, individuals evaluate these attributes in view of their personal goals and concerns to determine *relevance* and emotional *valence*. Finally, in the Responding phase,

based on the comparisons, individuals experience feelings of *involvement* and *distance* and may have *use intentions*, influencing overall *satisfaction*. Importantly, perceived *aesthetics* of and *similarity* with an agent moderate transition from Comparing to Responding. Understanding the agent's capabilities (*affordances*) and appreciating its appearance (*aesthetics*) may influence engagement decisions as well.

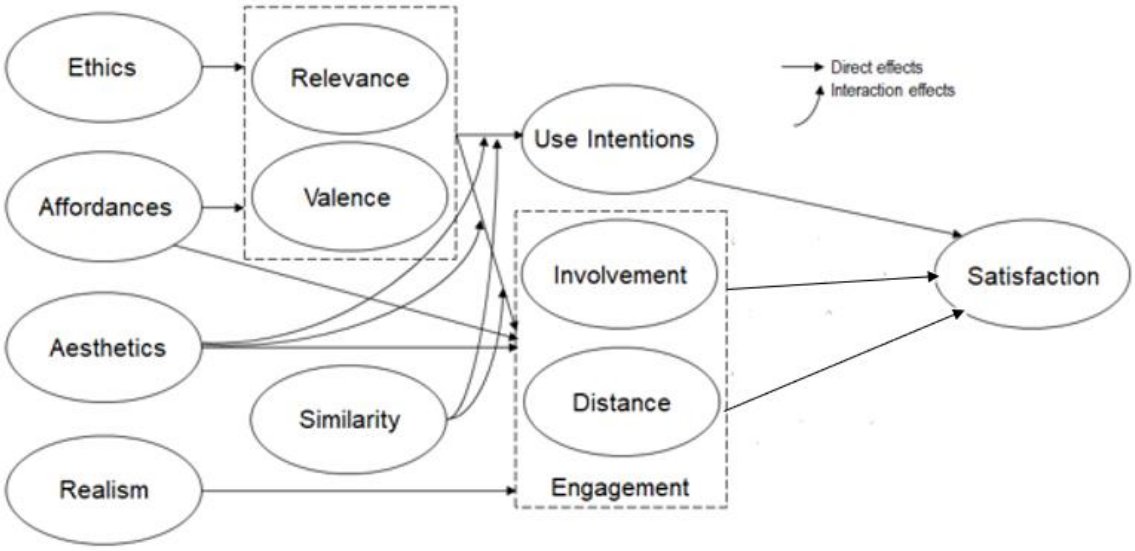

Figure 2. 7 Interactive Perceiving and Experiencing Fictional Characters (I-PEFiC) (Van Vugt, Konijn, & Hoorn, 2009)

The I-PEFiC model was developed to explain user engagement with relational technologies, such as social robots, by integrating concepts from several prominent technology adoption theories. While I-PEFiC is a distinct model, it does draw inspiration and key elements from established frameworks like TAM, UTAUT, TTF, and EDM.

The following sections will unpack how specific constructs and relationships from these prior models are reflected and adapted within the I-PEFiC framework. By highlighting these connections, we can better understand how I-PEFiC builds upon and extends the existing knowledge on technology adoption and user engagement.

2.3.1.4.1 How I-PEFiC relates to TAM3 (Affordances)

The I-PEFiC model's concept of affordances can be closely related to the TAM3. In TAM3, perceived usefulness and perceived ease-of-use are critical determinants of a user's intention to use technology. Perceived usefulness assesses the technology's relevance to a user's specific goals within a given context, while perceived ease-of-use evaluates how the characteristics of the technology facilitate goal achievement.

In the context of TAM3, the evaluation process entails gathering information about various dimensions that can aid in achieving goals, which form the basis for assessing form the basis for assessing perceived usefulness and ease-of-use. For instance, a user planning a trip to Hong Kong might judge the usefulness of a travel website's chatbot based on whether it provides services that help achieve this goal. Concurrently, the user might assess the chatbot's ease-of-use by evaluating the interaction, performance, interface and etc.

These dimensions, when specifically encoded, align with the concept of affordances in I-PEFiC. Relevance in I-PEFiC, akin to perceived usefulness and perceived ease-of-use in TAM3, address the user's main concerns when interacting with technology in consumer context. This relevance profile shapers user intention in I-PEFiC, similar to how perceived usefulness and perceived ease-of-use influence intentions in TAM3.

In conclusion, the concept of gaining use intentions through relevance induced by appropriate affordances in I-PEFiC has a notable correspondence with TAM3. The connection between affordances and relevance is supported by TAM3's evaluation process of perceived usefulness and perceived ease-of-use. Additionally, TAM3 informs a two-dimensional structure for relevance in I-PEFiC because of two main goals within the general consumers.

2.3.1.4.2 How relates to UTAUT2 (Affective processing)

UTAUT2's hedonic motivation, which leads to behavioral intention, exemplifies a subset of the affective processing in I-PEFiC. In UTAUT2, along with external determinants like social influence, price value, and facilitating conditions, and the long-term use determinant of habits, there are internal determinants that influence technology use intention. UTAUT2's performance expectancy aligns with perceived usefulness in TAM3, and effort expectancy is akin to perceived ease-of-use (Venkatesh et al.,2003). Unlike TAM3, which considers Perceived Enjoyment as influencing perceived ease-of-use, UTAUT2 posits hedonic motivation as an independent factor influencing behavioral intentions.

For example, a user might find educational robots to have limited usefulness and fair ease-of-use in improving learning outcomes. However, the fun aspect of using the robot to alleviate study pressure could result in a willingness to use it.

I-PEFiC expands upon the affective decision-making presented in UTAUT2. In I-PEFiC, relevance and valence are key constructs for use intentions, evaluated through goal comparison, which includes both cognitive and affective processing.

Specifically, valence results from affective processing, with users typically experiencing positive valence when task features meet desired goals or help avoid undesired outcomes. Thus, the positive valence associated with hedonic motivation in UTAUT2 is considered part of the affective processing in I-PEFiC's goal comparison. However, it's important to note that valence differs from hedonic motivation as it emphasizes the emotional polarity induced by goals and concerns, while hedonic motivation stems from the pursuit of pleasure.

2.3.1.4.3 How I-PEFiC relates to TTF (Relevance)

TTF posits that a fit between technology characteristics and task characteristics enhances use intentions. In the context of UTAUT2, incorporating TTF can positively influence performance expectancy. Similarly, I-PEFiC models this through its goal-comparison rule, where technology characteristics equate to features, and matching task characteristics align with user goals.

TTF, as applied to performance expectancy and effort expectancy, can lead to increased relevance and, consequently, use intentions. For example, a robot with a casual and friendly facial expression (technology characteristic) that tells jokes (task characteristic) may relieve a user's tension (goal), representing a TTF. This TTF can also lead to either hedonic motivation or valence, resulting in more positive valence and use intentions.

2.3.1.4.4 How I-PEFiC relates to EDM (Valence)

The I-PEFiC model and the Expectancy-Disconfirmation Model (EDM) share a fundamental connection through the concept of valence. I-PEFiC defines valence as the result of comparing goals with encoded features, a process based on user beliefs. Affordances that support goal achievement tend to yield positive valence.

Conversely, EDM explains how individuals form attitudes towards a product or service through the comparison of expectations and actual experiences. Disconfirmation arises from mismatch between experience and expectation, either outperformance or underperformance, leading to various levels of satisfaction that shape individuals' attitudes and behaviors. Confirmation from matching expectation or positive disconfirmation from outperforming technology results in user's satisfaction.

Satisfaction, as conceptualized in EDM, embodies a positive emotional state. Within I-PEFiC, this state is akin to experiencing positive. Projecting EDM in I-PEFiC, expectations

depicted in EDM can be equated to the user's goals that arise from their beliefs. The actual experience comes from features encoded in various dimensions such as affordances.

Thus, the confirmation of disconfirmation processes in EDM are analogous to the indicators of goal comparison in I-PEFiC. For instance, whether an affordance effectively facilitates goal achievement is a crucial aspect of this comparison. The emotion-evoking nature of EDM's process parallels the affective processing involved in I-PEFiC's goal comparison.

However, EDM encompasses a broader scope than just the valence-producing process through affordance and goal comparison. EDM also considers the emotional response elicited by other elements, such as aesthetics, which are compared against expectations. For example, a user might prefer a certain design aspect of a robot, and their aesthetics evaluation of this aspect could induce positive emotions that enhance their engagement, which in I-PEFiC would be covered by the path from aesthetics to engagement. This positive emotion, although distinct from valence in I-PEFiC – which is directly tied to goal comparison – still plays a role in shaping user engagement.

The key contribution of I-PEFiC is the explanation of the distinct perceptual processes when engaging with relational technologies like social robots, shaped by concurrently feeling involvement while maintaining emotional distance. This dynamic experiential perspective differs from pragmatic cognitive adoption models.

2.3.2 Model Comparison and Selection

Because it encompasses and elaborates on the factors of other technology-acceptance models, the I-PEFiC framework was selected as the theoretical model for this thesis to understand depressed young adults' experiences and perceptions of robot-delivered therapy, which determines their use intention and engagement with robots for **eiIBM** exercises. This

selection was based on considerations of applicability to the study context, affective communication strategies, expandable experiential factors, and bidirectional relationships between constructs capturing the progression of user perceptions.

2.3.2.1 Applicability to Current Study

TAM series are suited for explaining adoption of emerging information technologies, especially in organizational contexts (Davis, 1989; Venkatesh & Davis, 2000). However, their focus on perceived ease-of-use as one of the key mediator limit applicability to initial evaluations rather than continued technology use (Bhattacherjee, 2001). Therefore, TAM was applied in understanding the initial deployment of mHealth app (Mouloudj et al., 2023) and mobile payment system (Daştan & Gürler, 2016).

In contrast, UTAUT series expand the scope to consumer adoption of integration technologies (Venkatesh et al., 2003; Venkatesh et al., 2012; Paulo et al., 2018), e.g., chatbot in travel website (Phaosathianphan & Leelasantitham, 2019). Nonetheless, constructs like *Price Value*, *Habit*, Social *Influence* and *Facilitating Conditions* bear limited relevance for investigating user intentions and engagement with a novel and private robot-delivered therapy solution as examined in this study. Specifically, *Price Value* is inapplicable given the emerging nature of this robot service. *Habit* holds low significance for an inaccessible technology where users are yet to develop usage routines, as noted by Tamilmani et al. (2018). *Social influence* is negligible due to the private one-on-one therapy context. Finally, *Facilitating Conditions* are excluded as a theoretical focus since researchers facilitate the interactions. However, UTAUT2 includes some cognitive determinants directed to *Behavioral Intention*, making it a framework to understand the pre-use considerations of the technology.

Unlike the above technology-centric models, I-PEFiC originates from decades of research into user interactions with relational agents (e.g., Van Vugt et al., 2009; Hoorn, Baier, Van Maanen, & Wester, 2023). It explains adoption of socially oriented technologies designed to establish bonding, like the therapeutic robot employed in this thesis. Furthermore, I-PEFiC inherently accounts for self-related factors, beyond the environmental emphasis in the TAM series and UTAUT series. This aligns with the low-intensity behavioral intervention (i.e., one-on-one therapy) in the present context.

2.3.2.2 Affective Communication Strategies

Unlike traditional computer-mediated communication (CMC), which assumes technology is an informational tool, social robots incorporate human-like characteristics that elicit affective, social responses beyond just cognitive ones. As Hoorn (2020a) argues, individuals may perceive a robot as a mere medium of communication but also as a social actor when interacting with it. Consequently, users may experience both cognitive and affective responses, depending on whether reflective or affective processes dominate their perception and experience of the interaction. Reflective processes involve logical, analytical evaluations of the robot's functionality, which the diverse TAM and UTAUT versions cater to, while affective processes encompass emotional responses to the robots' human-like relational qualities, which is consisted in I-PEFiC.

A growing body of literature shows that whether users view a robot as a tool versus a social entity includes the effectiveness of robot-delivered intervention. These studies reveal that users who perceive robots as social actors exhibit greater emotional engagement, trusting attitudes, and therapeutic benefits compared to those who perceive them as merely functional tools (Li, 2015; Gursoy et al., 2019). This highlights the significance of affective factors in shaping user acceptance and outcomes with robot-mediated communication.

Technology adoption models, such as UTAUT2 predominantly focus on cognitive drivers of use intentions without fully considering affective variables. Although UTAUT2 and its extension with EDM include Hedonic Motivation and Satisfaction, these variables reflect pragmatic technology perceptions rather than interpersonal emotional experience with agents. While such models may be suitable for studies involving robots designed for rational tasks, they may not adequately capture the affective dimensions crucial in human-robot interaction.

In the present study, users engage in emotional communication with robots, such as encouragement and persuasion, and their acceptance and engagement largely depend on the affective processes arising from interpersonal bonding during therapeutic conversations. Therefore, I-PEFiC is a more appropriate framework than the Technology Acceptance Model (TAM) series and UTAUT series for this context.

2.3.2.3 Expandable Experiential Factors

Another advantage of I-PEFiC is its ability to incorporate distinctive features of various entities in technology, providing greater explanatory power for use intention and engagement. The robot-incorporated technology features varying affordances that could separately come from the robot or the exercise as distinct entities. For instance, a user might prefer the therapist but not be satisfied with the therapeutic approaches when therapist's companionship affordance was valued, but the therapy affordance of the therapist's approaches does not satisfy the patients.

The different affordances are perceived by users in terms of their respective relevance and valence, which shape the users' overall engagement and intentions. I-PEFiC's expandability enables tailored investigation of how distinct affordance influence user experiences across different contexts.

In contrast, UTAUT2's static set of determinants cannot account for the various affordances of integrated technologies like social robots. The predefined construct in UTAUT2 may neglect the relative significance of determinants across different technologies and contexts (Bagozzi, 2007).

Therefore, I-PEFiC's openness to modeling multifaceted experiential variables makes it better suited than UTAUT2 for understanding users' subjective perceptions and evolving intentions when interacting with a novel relational technology like robot-assisted therapy. The framework allows for defining variables based on the specific design and affordances, providing greater flexibility and adaptability compared to other technology adoption models.

2.3.2.4 Bidirectional Relationships

Finally, I-PEFiC captures the dynamic evolution of user experiences over time through bidirectional relationships between constructs (Hoorn, 2022b). It posits that initial expectations shape ongoing affordance perceptions, which in turn inform experiences and evolving intentions. I-PEFiC emphasizes expectation confirmation/disconfirmation processes underlying users' developing perceptions. In contrast, UTAUT2 only examines static predictive relationships between constructs, going into one direction. UTAUT does not account for how initial expectations may be modified by experiences to shape subsequent intentions and behavior (Venkatesh et al., 2012).

I-PEFiC's incorporation of bidirectional effects between variables provides a more comprehensive representation of real-world technology acceptance as a fluid process, continuously shaped by users' changing expectations and perceptions. This dynamic perspective is especially relevant in novel relational contexts like robot-assisted therapy, where users have limited initial expectations and develop affordance perceptions through experiential interactions over time.

With the above reasons, I-PEFiC is suitable to provide theoretical insight into depressed young adults' emerging perceptions of and sustained engagement with robot-delivered therapy. Next, I will introduce the application of I-PEFiC in understanding users experience of **eiIBM** delivered by social robots.

2.4 Research Framework of the Current Thesis

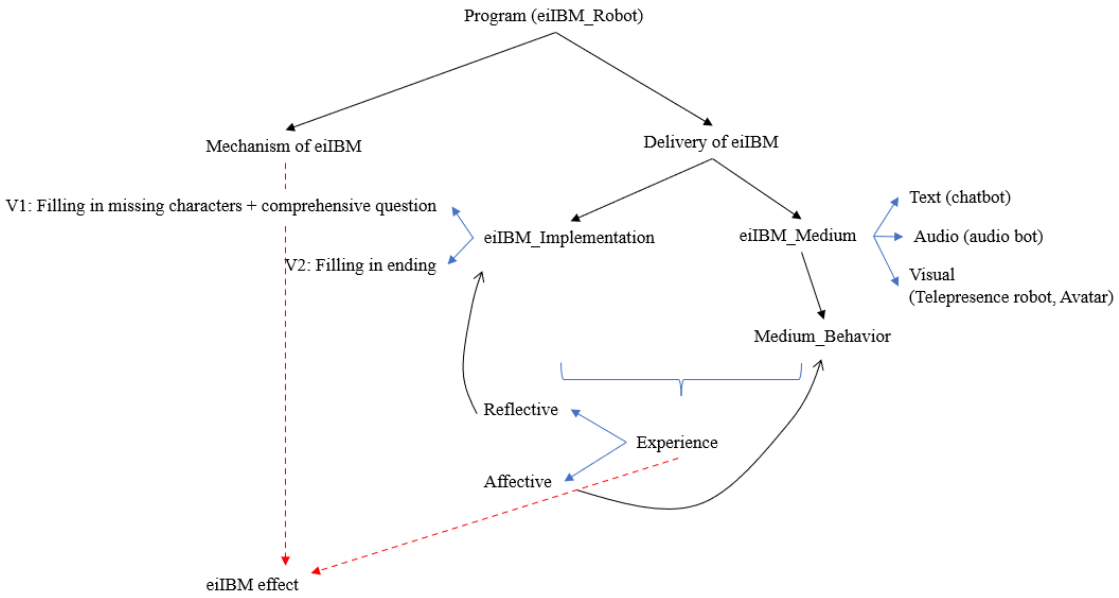

Figure 2. 8 The Components of Elaborative- and Imagery-based Interpretation Bias Modification Guided by Robot (eiIBM_Robot)

The research framework for this thesis integrates the elaborative- and imagery-based Interpretation Bias Modification (**eiIBM**) program with various robot modalities (**eiIBM_Robot**). The user experience is derived from the **eiIBM** exercise delivered by robots (*eiIBM_Delivery* component), which consists of the **eiIBM** exercise format (*eiIBM_Implementation*) and the robot with different modalities (*eiIBM_Medium*), and the robot's behavior (*Medium_Behavior*) (Figure 2.8). The robot's behavior is a function that depends on the robot and the **eiIBM** exercise format. Users are expected to gain both reflective and affective experiences through this integrated intervention. These experiences,

along with the empirically-supported mechanism of **eiIBM**, are hypothesized to contribute to the effects on cognitive bias reduction and depression alleviation.

The I-PEFiC model serves as the general framework for understanding affective user responses to artificial agents. The present study focuses on the affordances encoded from the features of the robot modalities, rather than the aesthetics or epistemics. The two main perceived affordances are “intervention delivery” and “ease-of-use interaction”. "Intervention delivery" is mostly derived from the **eiIBM** delivery, while "ease-of-use interaction" is mainly derived from the robot.

Considering the two separate entities (robot and exercise) in the **eiIBM** program, the present study first explores the varying robot effect with the standardized **eiIBM** exercise format, followed by an improved variant **eiIBM** exercise along with the most effective robots.

Study 1 maintains the evidence-based traditional **eiIBM** exercise format and employs chatbot, audio bot, and telepresence robot to deliver the **eiIBM** program (**eiIBM_RobotV1**). The objectives are to explore the effect of robot modalities on user experience and therapy effectiveness (**Objective 1** and **RQ1**).

Study 2, a complementary qualitative study using semi-structured interviews, further investigates depressed participants’ experiences and perceptions of **eiIBM_RobotV1** within the I-PEFiC framework. This study provides additional insights from a within-subject perspective on the different **eiIBM** robots, uncovering themes regarding how depressed users perceive and respond to distinct social agents for therapy (**Objective 2** and **RQ2**).

Based on the findings from Study 2, the **eiIBM** exercise format is modified to align with participants’ expectations and needs (**eiIBM_RobotV2**). Study 3 employs the visual modality in the form of virtual avatar which was preferred by depressed participants to

deliver **eiIBM_RobotV2**. The effects of the **eiIBM** medium, user experience, and intervention outcomes are re-examined (**Objective 3** and **RQ3**).

These studies contribute to the understanding of how user experience, influenced by various affective decision-making variables, impacts the outcomes of robot-delivered therapy. They also provide insights into how robots as CBT agents change participant expectations, informing the design of robot-augmented therapies.

Chapter 3: Research Design and Methodology

This chapter discusses the chosen research methodology to reach the research objectives, to be more specific, describing the selection, design and implementation of mixed methods guided by a theoretical framework.

3.1 General Methodology Description

3.1.1 Research Paradigm- 3rd Paradigm of Human-Computer Interaction (Harrison et al.,2007)

Based on the research objectives and plan outlined in Chapter 2, this work is grounded in the third paradigm of Human-Computer Interaction (HCI) as described by Harrison et al. (2007). The first paradigm of HCI, human factors, focuses on optimizing the fit between humans and machines to address specific usability issues. This pragmatic engineering approach does not align with the research goals here, which aim to understand subjective user experiences with relational technologies like social robots.

The second paradigm, classical cognitivism, models interaction as information processing and communication between computers and humans (Harrison et al., 2007). It emphasizes developing predictive models and generalizable principles for efficient system design. However, this study investigates the nuanced, situated experiences of users interacting with an emotional, relational technology. The second paradigm's emphasis on rational analysis and optimization cannot fully capture the complexity of human-robot relationships.

In contrast, the third paradigm views interaction as phenomenologically situated, with meaning constructed through experiences in context (Harrison et al., 2007). It focuses on supporting situated meaning-making and appropriation of technologies. This aligns closely with the research objectives here to understand depressed individuals' perceptions and

adoption of robot-assisted therapy using the I-PEFiC framework. The third paradigm allows incorporating multiple theoretical lenses to analyze the experiential, affective, and contextual factors shaping user adoption of social robots (Harrison et al., 2007). It values pluralistic interpretation over reductive models, accommodating the intricacy of real-world technology use (Harrison et al., 2007). This interpretive flexibility suits the exploratory nature of this research into a novel relational technology for mental health. Furthermore, the third paradigm recognizes knowledge as situated and experiential rather than abstract and universal (Harrison et al., 2007). This fits the elicitation of subjective user viewpoints through interview (in Study 2), expanding the interpretation of the variables in the I-PEFiC framework.

3.1.2 Pragmatic Mixed Methods Research Methodology

A mixed-methods approach combining qualitative and quantitative viewpoints, data collection, and analysis (Schoonenboom & Johnson, 2017) was employed in this study. It aligns with the pragmatic worldview in the third paradigm of HCI, emphasizing solving real-world problems using multiple approaches rather than methods alone (Tashakkori & Creswell, 2007). The mixed-methods approach offers an abduction-intersubjectivity-transferability approach (Morgan, 2007) where reasoning iterates between induction and deduction, integrating subjective and objective stances as practiced by researchers.

Deduction involves developing hypotheses based on theory then designing studies to test them, moving from the general to the specific (Trochim, 2006). Induction analyzes observations to generate new hypotheses and theories, moving from the specific to the general (Creswell & Plano Clark, 2007). Specifically, abduction refers to utilizing both deductive and inductive reasoning to understand the research questions in context (Morgan, 2007). Intersubjectivity highlights considering interpretations from both researchers and

participants as valid (Morgan, 2007). Transferability examines applying knowledge created through action and reflection to a new setting (Morgan, 2007) - in this case, **eiIBM_RobotV2**.

There is growing acceptance that mixed methods better address complex real-world problems by acknowledging context, recognizing particulars and patterns, developing contextualized understanding, and balancing neutrality with advocacy (Greene, 2008). Since the 1980s, mixed methods have gained acceptance in human-computer interaction for triangulation (Mackay & Fayard, 1997; Johnson, 2004) and slowly in mental health care (Fitzpatrick, 2021).

3.1.3 Theoretical Framework-Guided Mixed Methods Research Approach

Morgan (2007) argues that researchers should connect issues in epistemology to research design using methodology, rather than separating the understanding of knowledge from efforts to produce it. Morgan (2007) states that mixed methods research may have an overall "theoretical drive" guiding the integration of qualitative and quantitative components.

The thesis emphasizes the importance of theoretical frameworks in directing mixed methods, as Morgan (2007) advocates. I-PEFiC provides the theoretical lens, making the researcher aware that the **eiIBM_Robot** contains different features potentially affording different perceived functions (van Vugt et al., 2009). The goal is to understand how users experience and perceive different robots and exercise presentation, and whether these experiences influence intervention outcomes of **eiIBM_Robot** in negative interpretation biases and depressive severity.

Based on Chapter 2, the researchers conceptualized the **eiIBM_Robot** as containing two key elements –Mechanism of **eiIBM** (its static and dynamic success factors) and *eiIBM_Delivery* (including *eiIBM_Implementation* and *eiIBM_Medium*). I-PEFiC helps

recognize how *eiIBM_Medium, eiIBM_Implementation* and their interplay might influence *eiIBM_Robot* acceptance and engagement, thereby affecting outcomes. Therefore, I-PEFiC enables the researcher to first use deduction approaches to understanding the experience with robots (*eiIBM_Medium*) to evaluate the research hypotheses. It also guided the induction approach to deeply understand variable relationships and hidden insights under the testing. Finally, to transfer the understanding to an improved context, the researcher returned to deduction to explore the experience when matching the *eiIBM_Medium* to *eiIBM_Implementation*. The cognitive theory – the remainder of mechanism of **eiIBM** in their robot-version variant – enabled understanding the effect of the experience on intervention outcome.

3.2 Research Methods – Experiments and Semi-Structured Interview

Study 1 and Study 3 necessitate examining the effects of *eiIBM_Medium* (robots) on experience and intervention outcomes while controlling for *eiIBM_Implementation* and remaining other conditions, as well as validate the hypothesized relationships within the I-PEFiC framework. Thus, experimental investigations were utilized in Study 1 and Study 3, as they enable confirming causal relationships between factors in a controlled setting (Kirk, 2013; Fox and Denzin, 1979). For the inductive portion in Study 2, semi-structured interviews were adopted to elicit additional qualitative insights into nuanced aspects of the experience.

As preparatory work, relational investigations through cognitive assessments allowed discovering connections between depression levels and cognitive states, using empirical tools requiring translation and validation for Chinese users over time. Following this preparatory phase, experimental investigations in Study 1 and Study 3 provided the opportunity to explore fundamental case relationships between robot features, individual experiences, and

intervention outcomes. The Study 2 qualitative data also enabled triangulation for interpreting the quantitative results. Employing this combination of methods facilitated a comprehensive understanding of the phenomenon under study.

Specifically, Study 1 combined with Study 2 represents an explanatory sequential mixed methods design. This approach is characterized by initial quantitative data collection and analysis, followed by qualitative data gathering to help explain or build upon the quantitative findings (Schoonenboom & Johnson, 2017). Together, Study 2 complements Study 1 by controlling preference biases of robots and further elucidating potentially non-existent differences. Study 2 itself utilizes a concurrent parallel mixed methods design, gathering qualitative and quantitative data simultaneously and integrating them during the interpretation phase (Schoonenboom & Johnson, 2017; Morse & Niehaus, 2009). It revealed subtle experiential nuances during human-robot interactions that may be omitted by controlled experiments.

Study 1 and Study 2 provided inspiration regarding the relationship between *eiIBM_Medium* and *eiIBM_Implementation* to potentially improve experiences and intervention outcomes in Study 3. Study 3 quantitative data allowed systematically assessing the effect of *eiIBM_Implementation*-improved **eiIBM_Robot** (**eiIBM_RobotV2**) on experience and effect in a controlled manner.

Next, I elaborate on the specific quantitative and qualitative methods utilized.

3.2.1 Quantitative Method: Experiment

3.2.1.1 Online Experiments and Lab Experiments

In the context of human robot interaction, the three most frequently used experiment categories are online experiments, lab experiments and field experiments. Among them,

online experiments and lab experiments are more appropriate solutions than field experiments for decreasing the impact of confounding variables on the research.

Specifically, lab experiments enable control over conditions to collect high-quality data. But they may lack ecological validity simulating real-world therapy contexts, causing disconnects between preferences, intentions and actual experiences (Levitt & List, 2005).

Although online experiments present challenges such as environmental confounds, higher dropout rates compared to face-to-face experiments (O'Neil, Penrod, & Bornstein, 2003), and the risk of contamination without stringent proctoring, their popularity has surged recently in comparison to traditional laboratory experiments. This trend can be attributed to several benefits. First, online experiments enhance participant comfort (Salgado & Moscoso, 2003). Second, they reduce the costs associated with experimental manipulation (Reips, 2002). Lastly, they offer improved generalizability of the results to a broader population (Huber & Gajos, 2020; Peyton, et al., 2020; Reips, 2000).

In Study 1 and Study 2, online experiments were adopted for three considerations. First, depressed youth tend to be characterized as less energetic and socially anxious but accustomed to social media such as WhatsApp and WeChat. Second, the goal was to determine long-distance therapy, and a natural ecological setting benefits the collection of true experience data. Third, the live video/audio/text communication format enables researchers to observe participants' involvement to reduce environmental impacts.

In Study 3, a lab experiment was used because the *eiIBM_Implementation* required verbal interaction with robots improved by natural language processing (NLP), needing high network and media quality. Although sacrificing naturalness, a private lab environment ensured communication quality while making depressed participants feel safe.

3.2.1.2 Experiment Setting

3.2.1.2.1 Wizard-of-Oz approach

Another issue to consider when designing experiments is how to control the independent variables to create multiple experimental conditions (Kirk, 2013). In certain experiments, control of the independent variable is quite easy and straightforward. For instance, Huang et al. (2023) controlled the conditions by having the voice agent or the robot guide the meditation. In practice, however, all robot systems make errors in terms of social rules or functionality. At present, there is no way to make an error-free robot system.

A solution is the Wizard-of-Oz approach (Martin & Hanington, 2012). That is, researchers can have a human secretly acting as the robot, listening and responding as the robot. This illusion is maintained until debriefing at the end of the experiment. Thus, participants believe they interact with an autonomous system. This approach enables testing ideal applications not yet existent.

However, it has limitations. Humans also make errors in listening and responding. It becomes difficult to tightly control the independent variable (Feng & Sears, 2009; Li et al., 2006). Therefore, in Studies 1 and 2, the fully Wizard-of-Oz approach was used by pre-scripting responses when participants' reactions in **eiIBM_RobotV1** were predictable.

In Study 3, with the open-ended **eiIBM** exercise format, participants were allowed to fill out the endings themselves. It was impossible for a human wizard to prepare responses immediately, neither could the feedback be pre-scripted. Therefore, another approach adopted in Study 3 was developing an artificial intelligent system to assist the human wizard in responding more consistently (Li et al., 2006). Therefore, robots for **eiIBM_RobotV2** were equipped with natural language processing (NLP) intelligence for natural communication and

were responsible for gaining responses from participants and generating feedback. However, a human assisted to review the generated feedback before sending it out to avoid errors.

3.2.1.2.2 Between-group Design

In Study 1 and Study 3, a between-subjects design was adopted because it could prevent confounding factor effects of previous experience. It enabled the inspection of the single influence of different *eiIBM_Medium* with the same *eiIBM_Implementation* on experience and intervention outcome.

Between-group design is also called “between-subject design.” In a between-group design, each participant is only exposed to one experimental condition. From a statistical perspective, between-group design is a cleaner design. Since the participant is only exposed to one condition, the users do not learn from different task conditions. Therefore, it allows avoiding the learning effect on another condition.

However, the results might be subject to substantial impacts from individual differences: the difference between the multiple values that are expected to be observed can be buried in a high level of “noise” caused by individual differences, which leads to large sample size. Therefore, it is harder to detect significant differences and Type II errors are more likely to occur. As in the condition in this study, when comparing the performance of one group of participants against the performance of another group of participants, the robot preference or expectation could create variance and therefore lead to false rejection of the alternative hypothesis. To minimize the impact of noise, an attempt was made to meet the minimum sample size calculated by G-Power beforehand. Also, a random sampling distribution method with well controlled gender, language, age and depression level of the participants was used to reduce bias caused by. Study 2 was used as the data triangulation of Study 1 as explained beforehand.

3.2.1.2.3 Within-group Design

In Study 2, a within-group design experiment was adopted in the quantitative component, because this design enables the participants to evaluate their experience with each specific type of robot (*eiIBM_Medium*) in comparison to the other ones. As a result, it enabled the researcher to understand the differences at an individual level rather than only on the group level.

The advantage of using a within-group design (also called "within-subject design") is that it requires each participant to be exposed to multiple experimental conditions, while only needing one group for the entire experiment. Compared to a between-group design, researchers can compare the performance of the same participants under different conditions. Thus, the impact of individual differences is effectively isolated. Moreover, the expected difference can be observed with a relatively smaller sample size.

However, a limitation of the within-group design is the difficulty in controlling learning effects. In Study 2, the experience of interacting with a previous robot type and the order of exposure might influence participants' evaluation of a specific robot (c.f. Bradley, 1958; Morii et al., 2017). To minimize such anchoring or order effects, an attempt was made to randomize the order of exposure across participants (Charness et al., 2012).

3.2.1.2.4 Materials for Experiment - Social Robot

For the studies, a total of four social robots were utilized. In studies 1 and 2, participants interacted with a chatbot and an audio-based robot through WhatsApp, without the physical presence of the robots. These represented the robot using text and voice modalities, respectively. Regarding the telepresence robot, which represented the visual modality, participants interacted with a NAO Humanoid Robot model, renamed Zora, developed by SoftBank Robotics, via a phone screen. Figure 3.1 display Nao robot though

the WhatsApp video call. As the NAO robot lacked a built-in Cantonese language library, its audio output was adapted to Cantonese and Mandarin using Google's text-to-speech service.

Figure 3. 1 Nao robot on Phone Screen through WhatsApp Video Call

In study 3, an avatar-based robot replaced the on-screen NAO robot as the visual modality, based on insights from second studies in Chapter 6. Alongside the avatar-based robot on the screen, an audio-based robot was also present, but without a visual avatar. Both the avatar-based robot and the audio robot utilized gender-neutral appearances and voices to avoid any potential biases.

Figure 3. 2 Avatar-based Robot on Computer Screen

3.2.1.3 Sampling Strategy

The definition of young adults as those aged between 18 and 30 is based on developmental and sociological research. This age range is recognized as a distinct developmental period known as "emerging adulthood" (Arnet, 2000). During this period, individuals experience significant changes and challenges related to identity, independence, and responsibility. Arnet's research (2020) argues that these experiences are sufficiently distinct from both adolescence and later adulthood to be considered a separate developmental period. Moreover, psychological research has suggested that cognitive development continues into the mid-20s, further supporting the upper limit of 30 years (Giedd, 2004).

Purposive sampling is a non-probability sampling method that is characterized using judgement and a deliberate effort to obtain representative samples by including typical areas or groups in the research (Etikan, Musa, & Alkassim, 2016). In this study, purposive sampling is used to select individuals who meet specific criteria, namely, young adults in Hong Kong aged 18-30 who are experiencing emotional distress. This method is suitable for this study, as it allows the researcher to focus on a specific group of individuals whose experiences and perspectives are most relevant to the research questions.

3.2.2 Qualitative Method: Semi-Structured Interview

In Study 2, semi-structured interviews were conducted to obtain insight into the experiences of the participants that were omitted by the controlled experiments in Study 1. Each interview lasted 30 minutes, consisting of open-ended questions about participants' emotional distress, experience with the robot-assisted exercises, preferences across robot delivery mediums, and opinions on future applications.

The interviews aimed to establish contextual information on participants' emotional distress and mitigation attempts. Next, their experiences with different robot mediums were examined, including preferences and perceived differences. Willingness to engage in future online robot training was also explored, along with trust criteria for exercise robots.

The semi-structured format allowed for a natural flow while ensuring that all topics were covered. Supplementary questions were clarified and probed in detail. The interviews were recorded and transcribed for analysis.

Several factors dictated this qualitative approach. Semi-structured interviews are a proven technique for acquiring common understandings, opinions, attitudes and experiences from participants sharing a common phenomenon (Arksey & Knight, 1999), namely interacting with the **eiIBM_RobotV1** here.

The flexible structure further allowed for tailoring questions to individuals while minimizing stress, revealing personal interpretations of the experience (Miles & Hubermann, 1994). This technique was well-suited to eliciting additional experiential insights beyond the controlled experiments.

Overall, the semi-structured interviews enabled eliciting detailed participant perspectives on interacting with the robots and exercises, providing a rich qualitative complement to the quantitative findings of Study 1.

3.3 Data Collection

3.3.1 Data Collection Strategies

This work employed two complementary data collection strategies for assessing outcomes and experiences - pre-post testing to evaluate changes before and after the **eiIBM_Robot** intervention, and repeated measures to collect multiple measurements under different conditions for evaluating user experiences. The former was applied in Study 1 and 3, while the latter utilized both complete (Study 2) and between-subjects (Study 1 and 3) approaches. These strategies enabled assessment of both the intervention effects and subjective experiences.

3.3.1.1 Pre-post Testing

Pre-post testing refers to a research design where measurements are taken before (pre) and after (post) a treatment or intervention (Dimitrov & Rumrill, 2003; Christ, 2007). It provides a simple way to evaluate the effect of an intervention by comparing before and after measurements. The pre-post-test design is commonly used in clinical and therapy-related research to assess changes resulting from the intervention (Huang et al. 2023; Luo et al., 2022; Dimitrov & Rumrill, 2003).

In the context of this work, the pre-post testing design was adopted in both Study 1 and Study 3. In both the control group and the intervention groups, I used pre-post measurements of the depression-related outcome measures. Specifically, in Study 1 and Study 3, depression-related assessments were conducted at baseline and after the 2-week intervention period for both control and intervention participants. This design allowed observing whether **eiIBM_Robot** had any effect on the participants' depression-related assessment results compared to the control group.

3.3.1.2 Repeated Measures Design

In all three studies, the repeated measures design was employed. Repeated measures design refers to an experimental design in which multiple measurements of the same variables are taken on the same subjects under different conditions (Kraska & Marie, 2010). Repeated measures design can be divided into two main types: complete repeated measures design and between-subjects repeated measures design (Girden, 1992).

Complete repeated measures design requires each subject to experience all treatment conditions, which can reduce individual differences and improve statistical power (Kraska & Marie, 2010). Study 2 adopted a complete repeated measures design, where each participant interacted with different types of robots (text-based, voice-based, embodied) over one week involving **eiIBM_RobotV1**. This design allowed the researcher observing their experiences with different robots. Each robot was evaluated after each session to prevent participants from forgetting previous experiences.

Between-subjects repeated measures design uses different subject groups to represent different treatment conditions, with each group measured repeatedly on the same variables at different time points (Girden, 1992). Study 1 and Study 3 employed a between-subjects repeated measures design, with experience of *eiIBM_Medium* measured repeatedly across time. This design permitted comparing experience changes over time within each condition, and comparing cognitive assessment data between the two studies, thereby elucidating the effects of two versions of **eiIBM_Robot**. Compared to complete repeated measures, this design requires a larger sample size but enables comparisons between intervention groups.

3.3.2 Data Collection Techniques

There are four kinds of data collected through three data collection techniques.

3.3.2.1 Demographics

The research scope led to examining the context of participants' age, gender, language, and depressive severity as part of controlling the experience of **eiIBM_Robot** and its effect on depression-related outcomes. A demographic questionnaire was used to collect their age, gender and language data, and a depressive questionnaire to understand their depressive severity. Gender (Female/Male) and language (Mandarin/Cantonese) are nominal data, whereas age and depressive severity are continuous scale data. The demographics were collected in the registration phase.

3.3.2.2 Variable-Oriented Data

To test the hypotheses, data were collected on experiential variables derived from the I-PEFiC framework such as affordance to therapy delivery, the relevance and valence of this affordance, and the use intention and engagement with **eiIBM_Robot**. For each experiential variable, there were more than 4 but less than 6 statements (Hinkin et al., 1997) with a 6-point Likert-type rating and 1/3 of the statements were counter-indicative to reduce acquiescence bias (Jackson & Messick, 1965; Javeline, 1999). Likert-type statements with 6-point rating scales obtained the participant's degree of agreement with the statements about the experiential variables (i.e., 1 = "Strongly disagree", "Disagree", "Slightly disagree", "Slightly agree", "Agree", "Strongly agree" = 6). A six-point rating scale avoids neutral bias among the participants and forces them to choose a side: A 'neutral' position follows from averaging over indicative and contra-indicative items while the mean turns out to be scale midpoint. The variable-oriented data were collected in Study 1, 2 and 3.

3.3.2.3 Assessment data

The intervention outcome measures included the near-transfer index – elaborative negative interpretation bias tendency, the far-transfer index – depressive severity, and an

indirect index of negative bias – automatic negative interpretation bias. These data were collected through empirically evidenced assessment tools. The raw data included nominal data (i.e., False/True) or ordinal data (i.e., totally disagree, disagree, agree and totally disagree). The bias indexes were finally processed according to their own coding manual and converted into ratio data. The assessment data were collected in Study 1 and 3.

3.3.2.5 Textual data

Study 2 conducted semi-structured interviews to elicit participants' rich interpretation when interacting with different types of robots. The interviews were audio recorded and transcribed verbatim into textual data for analysis.

3.4 Data Analysis Techniques

3.4.1 Experimental Analysis

A range of statistical analysis techniques were employed to validate the data, understand variable relationships, assess group differences, and model path relationships to comprehensively address the research questions.

3.4.1.1 Main Techniques for Data Validation

For the experiential variables, internal reliability and construct validity of the scales were evaluated using Cronbach's alpha and principal component analysis (PCA) in SPSS. These analyses were conducted to ensure the scales consistently measured the intended constructs across participants. High Cronbach's alpha indicates homogeneity or internal consistency between scale items intended to measure the same construct (Streiner, 2003). PCA reveals the underlying factor structure, confirming that scale items group into factors representing the intended constructs (Costello & Osborne, 2005). In this study, the PCA used Promax rotation with Kaiser normalization and free fitting format. Extraction and retention of factors were based on visual examination of the scree plot (Cattel, 1966) and eigenvalues

of >1.0 (Kaiser, 1960). In PCA, a factor loading threshold of .50 was applied for small sample size (30-50; Samuels, 2017) to enhance the strength of factors, so only items explaining ≥ 25% of the variance were retained. Assessing reliability and validity ensures the scales reliably measure the intended latent constructs required for the study.

For the intervention outcomes, the Shapiro-Wilk test assessed normality to ensure suitability for subsequent parametric analyses. Chi-square tests and independent t-test (for normally distributed variables; Mann-Whitney U tests for non-normally distributed variables) were conducted to check whether random assignment of participants was successfully controlled. These tests allowed verifying the groups were equivalent at baseline. Outlier detection identified and removed extreme values to prevent them from distorting subsequent analyses.

3.4.1.2 Data preparation

For the experiential variables, means were calculated for the remaining items within the same scale, turning the ordinal data into interval data for further analysis. This transformation enabled using parametric tests that assume continuous data.

Additionally, hierarchical cluster analysis (HCA) grouped participants by homogeneous experience patterns, exploring sample hierarchical organization (Lee & Yang, 2009). HCA was used to identify subgroups of participants with similar experience profiles to examine differential effects. There are two main approaches to resolve the grouping problem in HCA, agglomerative or divisive. In agglomerative HCA, each sample starts as its own cluster, and cluster pairs are recursively merged. In divisive HCA, all samples start as one cluster that is recursively split. Clustering uses a distance metric (e.g., Euclidean distance) and linkage criterion. Common linkage criteria include complete, single, average, and Ward's linkage. Ward's method minimizes within-cluster variance (Miyamoto et al., 2015). Ward's

method with squared Euclidean distance was used to categorize participants based on experiential variables (interval data), ranking subjects into low, medium and high experience groups.

For the intervention outcomes, pre- and post-data were calculated separately. Residual change scores between pre- and post-data were also obtained. Residual change scores estimate the predicted post-test scores by regressing the post-test scores on the pre-test scores, ignoring group assignments and then subtracting the predicted post-test scores from the observed post-test scores (Kisbu-Sakarya et al., 2013). It isolates the amount of change controlling for initial pre-test levels providing a purer measure of change compared to raw difference scores (Jennings & Cribbie, 2021). Compared to directly differencing the raw scores, residual change scores have the following advantages: 1) less affected by the ceiling or floor effects, where difference scores might lose efficacy; 2) provide greater variance in change by controlling for the pre-test, whereas the variance of difference scores is often lower, resulting in weaker test efficacy; 3) reduce errors caused by regression to the mean, while difference scores cannot control for this trend (Jennings & Cribbie, 2021).

3.4.1.3 Techniques to Understand Variable Relationships

Pearson's correlation analysis was performed in SPSS between demographic and experiential variables. This determined whether any demographic covariates needed controlling in subsequent comparative analyses based on their relevance to the dependent variables (Mertler & Reinhart, 2016). Identifying correlated covariates to include prevents them from confounding group differences.

3.4.1.4 Techniques to Compare Groups on Experiential Variables

Generalized estimating equations (GEE) were implemented in SPSS to assess changes between timepoints and between robot modalities (*eiIBM_Medium*) on experiential variables.

GEE accounts for non-normal distribution, within-subject correlations, and between-subject differences suitable for the current data characteristics (Ballinger, 2004). GEE provides a flexible approach for modelling correlated data that violates traditional (multi)variate analysis of variance assumptions.

Bayesian analysis of variance and Bayesian nonparametric tests were performed in JASP to compare experiential variables between robot modalities. Bayesian methods quantify evidence for the alternative hypothesis and incorporate prior knowledge, appropriate for evaluating similarity versus dissimilarity hypotheses with limited sample size (van de Schoot et al., 2013). Pairwise comparisons were used to inspect the difference across robot modalities and timepoints.

3.4.1.5 Techniques to Compare Groups of Intervention Outcomes

GEE models in SPSS analyzed pre-post changes between robot modalities (*eiIBM_Medium*) on outcome indicators, determining intervention efficacy and comparing conditions (including a control group without any intervention). GEE was again to handle the non-normal, correlated in time span data structure. Differences were compared using residual change scores which isolate pure change while controlling for pre-test scores (Kisbu-Sakarya et al., 2013). Pairwise comparisons were used to inspect the difference across robot modalities.

3.4.1.6 Techniques to Model Path Relationships of Experiential Variables

Partial least squares structural equation modelling (PLS-SEM) was applied in SmartPLS to validate the I-PEFiC framework by testing hypothesized experiential variable paths. PLS-SEM was chosen because it suits small samples and non-normally distributed data and directly examines proposed theoretical relationships (Hair et al. 2018). PLS-SEM enables testing complex models with many constructs and relationships, making it ideal for validating

the multifaceted I-PEFiC framework (Van Vugt et al., 2009). Multi-group analysis (MGA) compared experiential variable paths across timepoints to assess relationship stability.

3.4.1.7 Techniques to Compare the Intervention Outcomes across Different Experiential Groups

One-way multivariate analysis of variance (MANOVA) assessed intervention outcome residual change score differences between experiential groups ranked by cluster analysis. MANOVA was used to simultaneously compare multiple outcome variables between experiential groups while controlling the familywise error rate. Post-hoc Bonferroni comparisons identified specific between-group differences. One-way MANOVA also examined individual differences experiencing varied robot modalities (*eiIBM_Medium*) in Study 2.

3.4.1.8 Techniques to Compare between Study 1 and Study 3

Two-way MANOVAs with robot modalities (*eiIBM_Medium*) and **eiIBM** exercise format (*eiIBM_Implementation*) as factors compared the initial depression levels or cognitive biases to determine if differences on these variables explained effects on therapy outcomes. The two-way MANOVAs allowed assessing main effects of medium and implementation, as well as their interaction, on baseline symptomatology. This ensured the groups were equivalent before the intervention to attribute post-intervention changes to the experimental conditions rather than pre-existing differences. Subsequent two-way MANOVAs were conducted to examine the experience and intervention outcomes to explain the effect of the improved **eiIBM** exercise format and robot modalities.

3.4.2 Codebook Thematic Analysis Method

According to Corbin and Strauss (2014), qualitative data analysis involves multiple coding cycles that progress from initial open coding to more focused, selective coding as themes emerge. In this study, thematic analysis was utilized to analyze the interview data by identifying, analyzing and reporting themes related to participants' experiences and perspectives interacting with different robot modalities (Vaismoradi, 2013). Thematic analyses were chosen as a flexible method for systematically identifying patterns of meaning across the dataset to gain insights into the robot modalities preference and experience.

For coding, each participant utterance in the transcripts was defined as one conversational turn. Utterances were coded holistically, with each utterance potentially assigned multiple codes.

Ideally, three independent coders would code all utterances and inter-rater reliability would be quantified using Cohen's kappa (McHugh, 2012). However, due to time constraints, only one coder (the author) coded the interview transcripts. The author developed an initial codebook by open coding a subset of transcripts, resulting in a hierarchical code structure with top-level categories. Open coding allowed identifying initial concepts and categories directly from the data.

This codebook was iteratively refined by applying it to additional transcripts using constant comparison to merge, separate or enrich codes and identify salient themes related to the research questions (Glaser & Strauss, 2017). The final codebook captured key perspectives on factors influencing robot preferences, needs, and experiences among young adults with depression to address assumptions from Study 1 and explain differential outcomes in Study 2. The final codebook provides a framework of users' salient viewpoints for interpreting the quantitative results and generating design insights.

3.5 Ethical Considerations

Prior to study commencement, ethical approval was obtained from the University (Appendix E), with strict adherence to the principles of non-maleficence, autonomy, and confidentiality. Eligible participants were provided with an information sheet and their written or digital consent was obtained after explaining the study purpose and procedure, ensuring informed participation.

It was anticipated that exposure to ambiguous scenarios during the experiment might trigger emotional responses in participants. If a participant experienced emotional distress, the experiment would be terminated, and the participant's well-being prioritized in determining continuation. This precaution aimed to minimize potential harm.

To protect participants' privacy, all questionnaire data and interview audio files were password-protected, and participants were assigned pseudonyms. Coded identifiers were used to protect participant identities, with names and code numbers stored separately to maintain confidentiality (Ho, 2017; Coughlan, Cronin & Ryan, 2007). All research data will be securely discarded three years after study completion, with any identifying documents shredded.

While the Wizard-of-Oz (Woz) technique offers benefits in human-robot interaction studies, it raises ethical concerns due to its inherent deception potential (Lazar et al. 2010). Participants may mistakenly believe they are interacting with an autonomous robot, potentially influencing their emotional responses and perceptions.

To address these concerns, several measures were implemented. First, the robots were designed to avoid requesting personal information or intentionally eliciting emotional responses, reducing manipulation risks. This aligns with the American Psychological

Association's (2017) ethical guidelines emphasizing transparency and honesty in research involving deception.

Furthermore, a compensation plan was established to mitigate any potential emotional or psychological harm. Participants were promised access to the fully developed online robotic service upon study completion, serving as restitution for their time and potential discomfort while allowing them to benefit from the research advancements.

Finally, participants were informed about the deception during debriefing, preserving study integrity while respecting participants' rights and well-being. Research found that effective debriefing seemed to eliminate negative effects perceived by the participants who felt they had been harmed (Smith & Richardson, 1983; Holmes, 1973). By implementing safeguards and following established guidelines, the study aimed to uphold research integrity while respecting participants' rights and well-being.

3.6 Summary of the Methodology Framework

In summary, to examine the effect of different robot modalities (*eiIBM_Medium*) and **eiIBM** exercise format (e*iIBM_Implementation)* on user experience and intervention outcomes as well as to comprehend the reasoning behind their (in)effectiveness, a comprehensive methodology framework was developed (Figure 3.1). This research is grounded in the third paradigm of HCI which emphasizes situated, experiential interaction aligned with examining user experience and engagement of robot-assisted therapy (Harrison et al., 2007).

The study employs a pragmatic mixed-methods approach combining qualitative and quantitative techniques guided by I-PEFiC framework. I-PEFiC provides a theoretical lens for understanding how users perceive and respond to relational technologies like social robots. The research involves abductive reasoning, iterating between deduction hypothesis

testing experiments and induction exploratory via interviews to integrate subjective and objective data (Morgan, 2007).

The methodology consists of three interconnected phases. Study 1 utilizes between-subjects experiments to test the effects of different robot modalities (*eiIBM_Medium*) on user experience and intervention outcomes using an initial version of the robot-assisted intervention (**eiIBM_RobotV1**). Study 2 employs within-subject comparisons and semi-structured interviews to explore insights omitted in Study 1 and gather detailed subjective perspectives on interacting with different robots. Insights from Studies 1 and 2 inform the development of an improved robot-assisted intervention (**eiIBM_RobotV2**). Study 3 then re-examines the effects of robot of robot modalities in experience and outcomes using **eiIBM_RobotV2** through between-subjects experiments.

Online experiments in Studies 1 and 2 and lab experiments in Study 3 enable systematic assessment of robots' impacts on user experience and therapeutic outcomes. The semi-structured interviews in Study 2 complement the quantitative findings by revealing nuanced subjective perspectives. Each study utilizes appropriate protocols, validated measurements, and advanced analysis techniques to ensure methodological rigor.

This multiphase, multimethod framework provides a systematic approach integrating subjective and objective techniques, guided by theory, to comprehensively examine the effects of varied robot modalities on user engagement, experience, and outcomes of robot-assisted therapy. The combination of controlled experiments and interviews enables a multidimensional understanding of the phenomena. By building on prior findings and refining the intervention across studies, the methodology is well-positioned to generate meaningful insights for optimizing the design and delivery of robot-assisted mental health interventions.

Table 3. 1 Flow Diagram of the Methodology for Exploring Depressed Young Adults' Experience and Perception of eiIBM_Robot and Its Effect on Intervention Outcomes

Research Paradigm	Third paradigm of Human-Computer Interaction (Harrison et al., 2007).		
Research methodology	Mixed Method design guided by I-PEFiC		
Research approach	Deductive (Study 1)– Inductive (Study 2) – Deductive (Study 3)		
Sampling design	Purposive sampling		
Research Methods	Study 1: Between-subject (*Medium*) experiment with **eiIBM_RobotV1**	Study 2: Within-subject (*Medium*) experiment with **eiIBM_RobotV1**, followed by semi-structured interview	Study 3: Between-subject (*Medium*) experiment with **eiIBM_RobotV2**
	An explanatory sequential mixed methods design to explicit the experience and efficiency of **eiIBM_RobotV1.**		
Research objectives	Given **eiIBM_RobotV1,** understand **t**he effect of robots (*eiIBM_Medium*) on experience (RQ1.1) and intervention outcome (RQ1.2); and the effect of experience on effect (RQ1.3).	Explore why do the difference/indifference of effect of robots (*eiIBM_Medium)* (RQ2).	Given **eiIBM_RobotV2,** understand **t**he effect of robots (*eiIBM_Medium*) on experience (RQ2.1) and intervention outcome (RQ2.2); and the effect of experience on effect (RQ2.3).
Data collection	Interval data from experience variables derived from I-PEFiC Nominal or ordinal data from intervention outcomes though negative interpretation bias screening tasks.	Interval data from experience variables derived from I-PEFiC Textual data from semi-structured interview	Same as Study 1 but had several items improved
Data Analysis (Statistical analysis)	Reliability and Validity: Cronbach's alpha and Principal component analysis (PCA) Normal distribution: Shapiro-Wilk test Outlier exploration: Boxplot Randon distribution: Chi-square test, independent t-test, Mann-Whitney U test Experiential grouping: Hierarchical cluster analysis (HCA) Intervention outcome difference: Residual change score Correlation between variables: Pearson's correlation Groups comparison for dissimilarity hypotheses: Generalized estimating equations (GEE) Groups comparison for Similarity hypotheses: Bayesian ANOVA and Bayesian nonparametric tests Casual-effect hypotheses: Partial least squares structural equation modeling (PLS-SEM) and Multi-group analysis (MGA) Within- for dissimilarity hypotheses: Generalized estimating equations (GEE) *eiIBM_Medium* comparison for dissimilarity hypotheses: One-way Multivariate analysis of variance (MANOVA)		
Data Analysis (Qualitative analysis)	Codebook Thematic analysis method		

Chapter 4: Test of Assessment Techniques

This chapter employs quantitative methods to evaluate the suitability of a Chinese-translated, mobile application version of the depression severity assessment - the Beck Depression Inventory Second Edition (BDI-II) and three negative interpretation bias assessments: the Word-Sentence Association Paradigm for Depression (WSAP-D), the Ambiguous Situation Paradigm – Similarity Rating Task (SRT), and the Scrambled Sentence Paradigm – Scrambled Sentence Task (SST). Assessment data was collected twice, with a 14-day interval, from young adults experiencing depression, without any intervening training. Rigorous reliability and validity analyses were conducted to ensure these assessments can effectively measure the effect of the robot-assisted interpretation bias modification program (IBM) in subsequent studies.

4.1 Introduction

This chapter revisits the concept of cognitive biases, where certain types of information are consistently favored for processing over others (Mathews & MacLeod, 2005; Gotlib & Joormann, 2010), as discussed in Chapter 2. These biases can manifest at various stages of information processing, including attention, interpretation and memory-based reasoning. Notably, individuals with depression tend to prioritize negative information (Mathews, Ridgeway, & Williamson, 1996) and exhibit negative interpretations of emotionally ambiguous information (Eysenck, Mogg, May, Richards, & Mathews, 1991). This negativity bias contributes to the formation of negative memories that perpetuate pessimistic reasoning about life events. Even when depressed individuals perceive positive stimuli, they often struggle to disengage from concurrent negative stimuli (Keller et al., 2019, Caseras et al., 2007, Sanchez et al., 2013), leading to persistent negative interpretations.

As emphasized in Chapter 2, modifying interpretation bias is crucial for disrupting the depressive circuit. Beck and Haigh (2014) distinguished between automatic processing of stimuli (<1500ms) and reflective processing (> 5000ms). In daily life, familiar information tends to be processed automatically using habitual ways (Moors & Houwer, 2006). Altering these automatic processes necessitates reflective and cognitive modifications. The present thesis proposes that an elaborative interpretation bias modification (eIBM) program could effectively reshape cognitive biases. Automatic Paradigm for Depression (WSAP-D; Beard & Amir, 2009), can accurately measure automatic interpretation biases (Cowden, Hindash & Rottenberg, 2015; 2017). The similarity Rating Task (SRT; Mathews & Mackintosh, 2000) assesses elaborative interpretation bias, while the Scrambled Sentences Task (SST; Wenzlaff & Bates, 1998) captures biases at an intermediate processing speed (1500-4000ms).

Although these assessment tools have demonstrated effectiveness in assessing depression-related negative interpretation biases across various countries (Würtz et al., 2022; Gonsalves et al., 2019), their efficacy in the Chinese context remains understudied (Smith et al., 2017). Translation of these measures may introduce cross-cultural cognitive biases, potentially affecting priming validity (Smith et al., 2017). Previous comparison studies conducted in Western language were mainly about the emotional Stroop task (e.g., Eilola, Havelka, & Sharma, 2007; Sutton, Altarriba, Gianico, & Basnight-Brown, 2007; Winskel, 2013) that align with the mechanism of WSAP but they could not be directly applicable to the current research context (Chinese). Moreover, the limited explorations of the reliability of Chinese versions of cognitive measures (e.g., Smith et al., 2017) did not examine their validity in assessing depression severity.

To address these gaps, the present study developed and evaluated Chinese version of the BDI-II, SRT, SST, and WSAP-D. Rather than comparing them to their English counterparts, the primary aims were to establish 1) the reliability of the translated cognitive

bias measures and 2) their validity in correlating with depression levels, using the validated Chinese version of the BDI-II (Wang et al., 2011a; Wang et al., 2011b; Wang et al., 2020; Zhu et al., 2018; Byrne et al., 2004) as the reference indicator. The measures were administered at two time points (Day 1 and Day 14) to align with the BDI-II's assessment window.

It is hypothesized that the cognitive bias measures and their indicators will demonstrate temporal stability (**H4.1**). The negative bias indices from the SST and WSAP-D are expected to correlate strongly with BDI-II scores at both timepoints (**H4.2**), while the SRT indices may not (**H4.3**). This is because with longer processing times, individuals can engage in reflective processing to meet social expectations, even if such interpretations are not habitual. Positive bias indicators may not negatively correlate with BDI-II scores (**H4.4**), as depressed individuals can initially perceived single positive stimuli but struggle to disengage from negative stimuli when both valences are present (Keller et al., 2019; Caseras et al., 2007; Sanchez et al., 2013).

This study lays the groundwork for subsequent investigations by validating effective measures and an accessible platform for evaluating the efficacy of a robot-delivered online IBM program. Establishing the psychometric properties of the Chinese versions of these tools is crucial for ensuring valid assessment of the invention's impact on cognitive biases and depressive symptoms in the target population.

4.2 Methods

4.2.1 Participants

Participants were eligible if they met the inclusion criteria detailed in Chapter 3. The author confirmed eligibility before assigning participants to the evaluation tasks. In total, 18 depressed young adults ($M_{(age)}$ = 23.67 $SD_{(age)}$ = 3.38; 14 Female and 4 Male, 11 Cantonese speakers and 7 Mandarin speakers) completed the study. Four additional participants dropped

out midway due to finding the workload larger than anticipated or not completing tasks on time.

4.2.2 Design and Procedure

All participants completed each cognitive bias task (SST, SRT, WSAP-D) and depression assessment on Day 1 (*T1*) and Day 14 (*T2*) using the mobile app **IBMTest@POLYUSD**. The app was downloaded from the App Store for iOS devices or the author's personal website for Android devices. Tasks were completed sequentially in the order of BDI-II, SST, SRT, and WSAP-D. Each task included clear audio and text instructions and a benign stimulus exercise before the main task. Task order was not expected to affect results as no feedback was provided and participants were unaware of criteria to please researchers.

Participants could rest between tasks but not stop midway through a specific task. If a task was left incomplete, participants had to restart from the beginning. Researchers monitored progress via a backend website, offering assistance when sought and sending WhatsApp reminders for extended response times.

4.2.3 Materials

4.2.3.1 IBMTest@POLYUSD- Translation and mobile digitalization.

The cognitive tasks (BDI-II, SST, SRT, WSAP-D) were translated into traditional Chinese for Hong Kong participants and/or simplified Chinese for Mainland China participants. Specifically, the BDI-II scale that has been translated into Mandarin (Wang et al. 2011a; Wang et al. 2011b) was used and transformed into Cantonese. The last item in BDI-II about sexual desire was removed to adapted to the young adults' situation, as only 13.7% of Hong Kong secondary school students (around age 18 or below) have sexual intercourse

experience (FPAHK, 2023). The Mandarin materials of SRT and SST were the same as the previous study (Yiend et al., 2019) and transformed into Cantonese version. There was no established Chinese-version of the WSAP-D; therefore, the author proactively took on the task of translating the traditional Chinese versions.

The author, proficient in both Cantonese and Mandarin, initially translated the tasks. Research assistants from Hong Kong and Mainland China evaluated the translation quality. Written Cantonese and Mandarin share grammar, vocabulary, and idioms, differing primarily in character sets (traditional vs simplified). Several Cantonese sentences/wordings were revised in accordance with the Hong Kong culture through careful discussion. The focus was on ensuring conceptual equivalence rather than mere literal equivalence.

All tasks were digitized and incorporated into the mobile app IBMTest@POLYUSD for both Android and iOS platforms.

4.2.3.2 Beck Depression Inventory – Second Edition (BDI-II)

The BDI-II is a validated 21-item self-report scale assessing depressive symptom severity over the past two weeks (Beck, Steer, & Brown, 1996). Chinese materials were adapted from Wang et al. (2011a; 2011b). Scores range from 0-63, with higher scores indicating more severe symptoms. The BDI-II has demonstrated high validity (α=.89) and reliability (r =.75) (Erford et al., 2016). The Chinese version achieved strong reliability (Cronbach's α=.94, split-half =0.91) and validity (r =0.69, $p < 0.001$) (Lu et al., 2002).

4.2.3.3 Word Sentence Association Paradigm for Depression (WSAP-D)

The WSAP-D is an experimental task based on semantic association processing measures from cognitive psychology, used to assess automatic interpretation biases (Cowden Hindash & Amir, 2012).

In the original E-prime version, each trial begins with a central fixation cross (+), replaced by an ambiguous sentence, e.g., "*A warm feeling spreads from your stomach to your chest*" *("一股暖意從你的腹部擴散到胸部")*, for 1000ms. The sentence is then replaced by a single negative, e.g., "*Illness*" (*"生病"*), or benign, e.g., "*Soup*" ("*湯*"), word. Participants indicate if the word and sentence are related by pressing the left (related) or right (unrelated) mouse button. The word remains until a response is made. The next trial begins immediately after the response. A 10-item practice task is provided to ensure understanding. Participants are instructed to respond as quickly as possible.

In the mobile app version for this study, rather than pressing buttons, participants indicated relatedness by pressing "√" or "×" buttons. Reaction times and endorsement rates were used to determine automatic interpretation biases. Milliseconds are used to assess reaction times. Commonly, reaction times of less than 200 milliseconds, greater than 5000 milliseconds, or larger than 2.5 standard deviations from the participant's mean were typically eliminated (Cowden Hindash et al., 2015). However, no specific cut-off was set in the app to account for individual processing speed differences (e.g., some depressed individuals reported functioning more slowly than before; Stahl,2002; Lam et al., 2014) and avoid affecting emotional state. Nevertheless, they were reminded to finish the task as quickly as they can.

The WSAP-D allows extracting two distinct indices of automatic interpretative bias based on differential endorsement of positive and negative words associated with the same sentences. Endorsing negative words does not necessarily preclude endorsing positive words for the same sentence, highlighting the need to account for individual differences in interpreting the same materials. Based on this rationale, two sets of the Chinese WSAT were designed, each with 30 sentences. The materials for this task were adapted from Wenzlaff and

Bates (1998). In one set, sentences were followed by a negative or benign word; in the other set, the same sentences were followed by words of opposite emotional polarity. Uniformity of sentences across sets, contrasted with differing stimuli, helps offset some interpretative bias from individual understanding. Each participant completed 60 total sentences for comprehensive analysis of interpretative bias tendencies.

4.2.3.4 Scrambled sentence task (SST)

The SST measures interpretation bias by manipulating the emotional valence of words in scrambled sentences that participants unknowingly reconstruct to reveal their tendencies toward negative or positive interpretations (Wenzlaff & Bates, 1998; Wenzlaff, 1988, 1993).

Materials were adapted from Wenzlaff and Bates (1998) and Yiend et al. (2019), with 15 items for assessment. Each item consisted of six words, five of which could be unscrambled to form a grammatically correct sentence, either positively or negatively valenced based on word selection. Participants used five words to create each sentence as a statement, not a question. Though written Chinese (either simplified or traditional) typically includes small spaces between characters without larger spaces between "words", characters were grouped into clusters of no more than three characters to form "words" for task congruency with English (Smith et al., 2017). In the app, participants selected five words per sentence by tapping, displaying the selection order. Instructions were provided in text and audio, with a benign word exercise before the main task. An example is "Has Green Child The Eyes Blue" (有 綠色 孩子 這個 眼睛 藍色).

Participants unscrambled sentences under a 3-minute time limit (~ 12 seconds/item). They were encouraged to proceed if mistakes were made. Before unscrambling, participants also learned a 6-digit number to be recalled after the task, occurring twice (Day 1 and 14)

with different numbers (52430, 53261). Number recall required two consecutive correct attempts.

Responses were coded as: 1) unscrambled negative sentence (***SST_TN***), 2) unscrambled positive sentence (***SST_TP***), or 3) incomplete sentence (***SST_F***). Chinese is a pragmatically oriented language than relies more heavily on contextual cues than the grammatically structured English. Most Chinese words lack grammatical inflections, and meaning is derived mire from pragmatic context than rigid syntactic structures (Yeh, 2004). Due to this, the SST in English can preset a more "standard" sentence for comparison with participants' reconfigured versions. However, in Chinese, the same set of words can generate multiple semantically coherent sentence, making it difficult to predefine a singular "correct answer". For example: me to is life cruel good (我來說 對於 是 生活 殘酷的 美好的). In English, the grammatically sentence is 'life to me is cruel/good', whereas in Chinese, both '對我來說生活是殘酷的/美好的' and '生活對我來說是殘酷的/美好的' are valid sentences. Sometimes, participants would omit certain components due to the lack of subject-verb distinction in Chinese, resulting in sentences with only four words. I still deemed these as complete sentences. For example: who I dislike I am like (我 我 討厭 自 己 喜歡). In English, it is 'I dislike/like who I am', whereas in Chinese, both '我討厭/喜歡我自己' and '我討厭/喜歡自己' are valid sentences. Therefore, sentences were coded as positive if the overall meaning was positive, regardless of syntactic similarity to the standard English answer.

4.2.3.5 Similarity Ratings Task (SRT)

The SRT measures elaborative interpretation bias using emotionally ambiguous passages (Mathews & Mackintosh, 2000). The task requires participants to rate the similarity

of disambiguated statements to the original passages. Those with negative interpretation bias perceive negatively disambiguated statements as more similar to original passages than positively disambiguated ones. The SRT detects interpretation biases across disorders and vulnerabilities, including depression (Yiend, et al., 2014; Yiend et al., 2019).

Materials were customized from Yiend et al. (2019), with 15 ambiguous passages suitable for assessment. Passages were presented sequentially, with a button press for the next sentence. The final word of each. The final word of each passage was incomplete, requiring participants to fill in the missing letters. A comprehension question followed each passage. For example:

Slide 1 [title]: PRESENTATION

Slide 2 [sentence 1]: *You give a presentation during class*

Slide 3 [sentence 2]: P*eople look interested and applaud at the end.*

Slide 4 [sentence 3]: *However, you feel you cannot answer the last qu_s_i_n*

Slide 5 [question]: *Did you give a presentation during class?*

In Chinese, the final word's last character was removed as incomplete characters are impossible. Participants wrote the characters in a canvas at the bottom of the app and submitted it.

Traditionally, after all 15 emotionally ambiguous passages, participants rated the similarity of four sentences to each passage. They saw the title of each passage, along with four separate sentences: two sentences (target items) were linked to the ambiguous passage and offered disambiguated interpretations (positive or negative) of the previous passage; the other two sentences (foil items) were unrelated to the ambiguous passage but offered positively or negatively valenced interpretations of the emotional content. The perceived similarity was rated on a 4-point Likert scale, with 0-3 indicating "very different", "fairly different", "fairly similar", and "very similar", respectively. The target (positive and negative:

T+ and T- respectively) and Foil (positive and negative: F+ and F- respectively) sentences for the earlier example follow:

Your presentation is successful [T+]

Your presentation is unsuccessful [T−]

You are generally a good writer [F+]

You are generally a bad writer [F−]

The app version made two changes to reduce fatigue and confusion. The task was divided into sections of five passages each to prevent fatigue and familiarity bias from presenting all 60 sentences (4 per passage) at once. Additionally, the 0-3 scale was replaced with similarity explanations on four separate button options to avoid confusion with number meanings.

4.3 Data Analysis and Results

4.3.1 Analysis Plan

Data were analyzed using SPSS 28, with the significance level of $p < .05$. After coding and computing task scores, outliers in each assessment task were examined. Outliers could result from participants finding tasks difficult or responding in a biased manner. Two types of biases were of particular concern: 1) participants not taking the task seriously (possibly due to task difficulty or lack of motivation), referred to as "non-serious participants"; and 2) participants responding in a biased way to please others (e.g., affirming all statements with a positive meaning), referred to as "low effect participants."

Internal reliability and split-half reliability of each task were analyzed independently based on timepoints. Internal reliability was assessed by calculating Cronbach's α for bias indicator or trait scores across tasks. A Cronbach's α value of .7 or higher is commonly deemed sufficient for a psychological test (Kline, 1999). Split-half reliability was evaluated

using the Spearman-Brown formula with the sum of the odd and even items based on the indicators. Test-retest reliability was assessed using Pearson's correlation formula to test tasks stability over time. This analysis evaluated the correlation between bias indicators or trait sum scores at two separate time points per task. During this phase, participants exhibiting a significant increase or decrease in depression levels were omitted to prevent substantial fluctuation and response bias between timepoints.

To affirm the validity of bias tasks in diagnosing depression, Pearson correlations were executed between bias scores and depression symptom levels within each timepoint. Independent sample *t*-tests were employed to scrutinize bias scores difference between language of each task.

4.3.2 Reliability Analysis

4.3.2.1 BDI-II

Depression severity was scored by summing ratings for the 21 items, each representing a typical depressive symptom rated on a 4-point scale rating from 0 to 3. Sums were calculated as *DS1* and *DS2* for *T1* and *T2*, respectively. Due to technical issues, 4 participants (C_4, C_5, C_6, C_9) lost 1 or 2 missing items under poor network. Missing values were handled using mean substitution from other items at each timepoint. Resulting scores were denoted as *DS_MS_1* and *DS_MS_2*.

Cronbach's α and Split-half reliability were not required as BDI-II items independently describe symptoms. It is expected that depressed individuals exhibit distinct symptoms in a limited sample (Rohrbacher & Reinecke, 2014). The maximum total score is 63. Outlier exploration indicated participant C_18 had steep drop from severe to mild depression (41 to 19), an extreme change. Table 4.1 reports means and standard deviations of depression scores. Test-retest analyses were performed with and without the outliers. With n

= 18, the Pearson correlation coefficient between *T1* and *T2* were.588 (p = .010). Excluding the outlier (n = 17), the correlation improved to .772 (p <.001).

Table 4. 1 Reliability Analyses on Assessment Indicators

Task	Condition means	T1			T2			Test-retest reliability
		Mean (SD)	Cronbach's α	Split-half reliability	Mean (SD)	Cronbach's α	Split-half reliability	
BDI-II	DS_MS (n = 18)	28.00 (6.91)	/	/	24.28 (7.47)	/	/	r = .588 (p = .010)
	without outliers (n = 17)	27.24 (6.29)	/	/	24.59 (7.58)	/	/	r = .772 (p <.001)
WSAP	WSAP_BER (n =18)	.475 (.105)	.544	*rho* = .029 (.910)	.543 (.186)	.639	*rho* = .277 (.265)	r = .453 (p = .059)
	without outliers (n = 16)	.482 (.109)	.577	*rho* = .188 (.485)	.567 (.169)	.640	*rho* = .460 (.073)	r = .527 (p = .036)
	WSAP_NER (n =18)	.661 (.148)	.669	*rho* = .508 (.031)	.624 (.209)	.878	*rho* = .834 (<.001)	r = .889 (p < .001)
	without outliers (n = 16)	.659 (.144)	.654	*rho* = .473 (.064)	.613 (.200)	.873	*rho* = .817 (<.001)	r = .873 (p < .001)
SST	SST_TNR (n =18)	.628 (.227)	/	*rho* = .741 (<.001)	.565 (.232)	/	*rho* = .537 (.021)	r = .793 (p < .001)
	without outliers (n = 17)	.652 (.210)	/	*rho* = .696 (002)	.567 (.239)	/	*rho* = .569 (.017)	r = .872 (p < .001)
SRT	SRT_PT (n =18)	23.76 (6.01)	.665	*rho* =.610 (.007)	23.54 (7.14)	.824	*rho* = .578 (.012)	r = .759 (p < .001)
	without outliers (n = 16)	24.17 (6.27)	.691	*rho* =.590 (.016)	24.04 (7.38)	.818	*rho* = .531 (.034)	r = .759 (p < .001)
	SRT_PF (n =18)	16.41 (5.19)	.504	*rho* =.679 (.002)	16.72 (6.12)	.689	*rho* =.437 (.070)	r = .524 (p =.026)
	without outliers (n = 16)	16.27 (5.43)	.511	*rho* =.688 (.003)	16.63 (6.49)	.727	*rho* =.601 (.014)	r = .518 (p =.040)
	SRT_NT (n =18)	21.10 (4.89)	.609	*rho* =.119 (.638)	20.40 (5.91)	.772	*rho* = .605 (.008)	r = .628 (p =.005)
	without outliers (n = 16)	21.02 (5.19)	.627	*rho* =-.057 (.833)	19.92 (6.08)	.772	*rho* =.684 (.003)	r = .645 (p =.007)
	SRT_NF (n =18)	13.78 (6.35)	.849	*rho* =.781 (<.001)	13.79 (6.40)	.819	*rho* =.480 (.044)	r = .689 (p =.002)
	without outliers (n = 16)	13.37 (6.39)	.837	*rho* =.749 (<.001)	13.32 (6.65)	.823	*rho* =.629 (.009)	r = .707 (p =.002)

4.3.2.2 WSAP-D

Coding and Scoring. Following previous studies (Cowden Hindash & Amir, 2012; Cowden Hindash & Rottenberg, 2015), responses were classified into four categories: 1) Endorsement of Negative Interpretations (*WSAP_NE*), 2) Rejection of Negative Interpretations (*WSAP_NR*), 3) Endorsement of Benign Interpretations (*WSAP_BE*), and 4) Rejection of Benign Interpretations (*WSAP_BR*). Score ratios were calculated for each category: *WSAP_NER*, *WSAP_NRR*, *WSAP_BER* and *WSAP_BRR*. For example, *WSAP_NER* was calculated by dividing *WSAP_NE* by the total number of negative stimuli sentences (*WSAP_NE* + *WSAP_NR*). Value ranges from 0 to 1, with 1 representing the most severe state. A propensity to endorse negative interpretations over rejecting them (*WSAP_NER* > *WSAP_BER*) indicates a negative bias (Beard & Amir, 2009)

Outliers Screening. WSAP-D requires quick responses, making it prone to random responding. Therefore, random responding was tested by analyzing correlations between negative misses and positive endorsements across two stimulus sets. Set 1 contained 30 sentences with either 15 benign or 15 negative stimuli. Set 2 contained the same 30 statements with reversed polarity.

Participants may have felt both the negative and positive stimuli were associated with the same statement across sets, rating both as correct. This could not definitively be attributed to random responding. However, bias towards positive/negative stimuli should occur across both sets. Hence, non-serious participants were identified by analyzing value distances between negative misses and positive endorsements across set. WSAP-D was also likely to elicit intentionally polarized responses (e.g., endorse all the positive stimuli without considering statements). However, no evidence existed to screen and luckily no such participants were identified.

Participants C_11 and C_4 were flagged as potential non-serious responders. C_11 agreed with both positive and negative stimuli in Set 1 but disagreed more in Set 2. C_4 largely agreed with both in Set 1 but agreed with negatives and disagreed with positives in Set 2, suggesting a tendency to endorse negatives rather than just disagree.

Reliability analysis. With n = 18 (Table 4.1), Cronbach's α for sentences paired with positive words and negative words at *T1* and *T2* ranged from .544 and .878. Notably, Cronbach's α for negative words consistently neared or exceeded.70, demonstrating high internal consistency. Split-half reliability showed stronger, significant correlations between scores for two half negative word-sentences (*T1*: *rho* = .508, p = 031; *T2*: *rho* = .834, p < 001). In contrast, correlations for positive word sentences were non-significant (*T1*: *rho* = .029, p = .910; *T2*: *rho* = .277, p = 265). Test-retest reliability of WSAP positive score and WSAP negative score across timepoints demonstrated stronger and significant association

(*WSAP_BER*: r = .453, p = .059; *WSAP_NER*: r = .889, p <.001). Missing values were replayed by mean values within indicator.

Excluding outliers C_11 and C_4 altered significance patterns. WSAP_BER test-retest reliability became significant (r = .527, p =.036, n =16) while WSAP_NER Split-half reliability at *T1* became non-significant (*rho* = .473, p = 064).

Results indicate the measure has stronger reliability for assessing negative versus positive biases. High internal consistency and strong test-retest reliability suggest potential for evaluating negative biases, though internal reliability weaken when precluding outliers. The measure may be ineffective for assessing positive biases, as indicated by the non-significant or mildly significant correlations with low coefficients. This distinction may reflect positive wors sentences serving as the counter indicator of negative bias.

4.3.2.3 SST

Coding and Scoring. Studies employing the SST have typically prioritized split-half reliability (e.g., O'Connor et al., 2021; Würtz et al., 2022) over internal consistency, often measured through Cronbach's α. Cronbach's α may be less suitable for the SST given the binary nature of item coding (0 or 1). The Kuder-Richardson method could be a more appropriate alternative. However, the SST tends to generate more 'missing' data compared to self-report questionnaires, as participants may form grammatically incorrect or incomplete sentences within timed tasks. Although the number of such items was minimal in the current samples, individual items might have more 'missing' values than standard questionnaires or the must-answer WSAP-D. This can complicate internal consistency computation, particularly with case wise deletion methods. In my dataset, only four participants completed unscrambled sentences validly for Cronbach's α analyses. Therefore, I restricted my analysis

to split-half reliability. Participants who correctly completed four negative sentences out of ten differed in bias tendency from those who completed four out of six.

Building on previous research findings (Hirsch et al., 2018; Hirsch et al., 2020; Krahé et al., 2019), I computed a 'negativity index' (*SST_TNR*) for each participant. Sentences unscrambled into negative or positive forms were coded as 0 or 1, respectively. SST_TNR was calculated by dividing the total count of positively unscrambled sentences by the total number of grammatically correct unscrambled sentences per participant. This ratio measure ranges from 0 to 1, with 1 suggesting all sentences were unscrambled negatively.

Outliers Screening. For the SST, lower completion rates were considered possible signs of random responding. If more than 5 sentences were ungrammatical, the participant was assumed to be unfamiliar with or had difficulty with the task. According to the data screening, C_13 exceeded the limit (6 invalid response in *T1*) and was assessed as the non-serious participant. The SST task was unlikely to have low effect participants since it required relatively quick responses and press on the screen could not be withdrawn, allowing more accurate judgement of respond processes.

Reliability analysis. Across all participants (n = 18), split-half reliability (odd versus even items), reflected in the Spearman-Brown Prophesy Reliability Estimate, was strong for *T1* (*rho* = .741, $p < .001$) and weak for *T2* (*rho* = .537, $p = .021$). Including outlier C_13 decrease significance and the estimate at *T1* (*rho* = .696, $p = .002$) but increased them at *T2* (*rho* = .569, $p = .017$). Although internal consistency was not consistently strong across time, this does not imply the translated SST was unreliable. The odd-even split method was incidental, and items might have different difficulty levels after Chinese translation. Therefore, I focused on the reliability of the full SST. Test-retest analysis showed high stability over time, with correlation of *SST_TNR* between the timepoints of $r = .793$ ($p < .001$, n =18) and $r = .772$ ($p < .001$, n =17).

4.3.2.4 SRT

Coding and Scoring. I calculated the score for four sentence types: 1) Positive Target (*SRT_PT*), 2) Negative Target (*SRT_NT*), 3) Positive Foil (*SRT_PF*), and 4) Negative Foil (*SRT_NF*). Missing values were handled using mean substitution from other items at each timepoint.

Outliers Screening. In the SRT, particular attention was paid to possible response biases. Non-serious and low effect participants were identified in two ways. First, non-serious participants were identified by checking if they correctly answered scenario comprehension questions, indicating seriousness about the scenarios. If participants answered incorrectly for more than 5 out of 15 items, the reliability of their responses to subsequent statements was doubtful.

Second, outliers were analyzed based on score ratios for the four descriptions (PT: positive target, PF: positive foil, NT: negative target and NF: negative foil). The maximum score for each description type was 45 (15 items × 3 points). A midpoint standard of 22.5 was used to judge the agreement level. Non-serious participants were identified by observing undifferentiated agreement or disagreement across all description types for most scenarios (undifferentiated agree/disagree tendency). Low effect participants were identified by observing undifferentiated positive or negative agreement tendencies. The specific logic was as follows:

When NF/NT ≥ 0.8, the participant was considered not to differentiate between negative statements (foil or target). On this basis:

1. If NF and NT scores were low, but PF and PT scores approached 1 and were high, the participant was suspected of intentionally giving high scores to all positive statements without considering the context (bias code 1).

2. If NF, NT, PF and PT scores were all low, this indicated an undifferentiated disagreement tendency, termed unbiased rejection without considering context (bias code 2).
3. If NF, NT, PF and PT scores were all high, this suggested undifferentiated agreement, or unbiased acceptance without considering context (bias code 3).
4. f NF and NT scores were high, but PF and PT scores approached 1 and were low, the participant was suspected of intentionally giving high ratings to all negative statements without considering context (negative bias tendency without considering context; bias code 4).

Potential low effect participants (biased response) included C_12 and C14 with bias 1, and C_4 and C_16 with bias 2. Bias code 1 participants could not definitively be identified as outliers, as they may have associated positive foils with the context and rated them highly. Bias code 4 also could not be ruled out as their genuine rating tendency. However, bias codes 2 and 4 suggest more random or undiscriminating rating patterns, raising higher suspicion that C_4 and C16 were outliers based on their patterns.

Reliability analysis. Cronbach's α and Split-half reliability were performed on the scores for all four sentence types, both with and without outliers. As shown in Table 4.1, with all participants (n =18), split-half reliability for *SRT_NT* was adequate at *T2* (*rho* =.605, p =.008) but not *T1* (*rho* =.119, p =.638), remaining inadequate when excluding two depressive outliers. Cronbach's α indicated acceptable internal consistency for *SRT_NT* only at *T2*. Test-retest reliability was also weak (n = 18: r =.628, p =.005; n = 16: r =.645, p =.007). In contrast, *SRT_PT* demonstrated medium split-half reliability at both timepoints, strong test-retest correlation, and acceptable good internal consistency (medium Cronbach's α for *T1* and high at *T2*). These results suggest that in the SRT, endorsements for positive target sentences

serve as a more robust counter-indicator of reliability and stability in measuring negative bias.

4.3.3 Validity of Negative Interpretive Bias Measures

To assess the validity of the WSAP, SST and SRT measures in capturing interpretation biases related to depression, Pearson correlations were conducted between the depression severity scores (***DS_MS_1*** and ***DS_MS_2***) and each task's indices of negative interpretation bias. Pearson correlations were selected as the appropriate validity analysis to relate interpretation bias scores to depression severity across timepoints, without assumptions of causality (Table 4.2). Regression analysis was not used due to insufficient evidence that negative interpretation bias alone explained depression severity. Paired sample *t*-tests were also not applicable, as they lack modelling parameters to account for repeated learning effects.

4.3.3.1 WSAP-D

Endorsement rates for negative (*WSAP_NER*) and positive (*WSAP_BER*) words associated with ambiguous sentences were used to assess negative interpretation bias on WSAP-D task. *WSAP_NER* showed a moderate positive correlation with depression severity at *T1* ($r = .646$, $p = .004$) and a strong positive correlation at *T2* ($r = .773$, $p < .001$). In contrast, WSAP_BER exhibited negligible correlations with depression at both timepoints. These results indicate that WSAP_NER serves as a valid indicator of negative interpretation bias, demonstrating the expected significant positive relationship with depression severity.

Table 4. 2 Pearson Correlation Between Depression (BDI-II Scale) and Negative Bias Indicators (WSAP, SST, SRT) to Test Validity, With Outliers

Task ($N = 18$)	WSAP_NER	WSAP BER	SST_TNR	SRT_NT	SRT_PT	SRT_NT/ SRT_PT
T1	$r = .646$ ($p = .004$)	$r = .063$ ($p = .805$)	$r = .529$ ($p = .024$)	$r = -.137$ ($p = .587$)	$r = -.044$ ($p = .864$)	$r = .011$ ($p = .966$)
T2	$r = .773$ ($p < .001$)	$r = .233$ ($p = .351$)	$r = .563$ ($p = .015$)	$r = .209$ ($p = .405$)	$r = -.591$ ($p = .010$)	$r = .495$ ($p = .037$)

Note. WSAP_NER represents the endorsement rate for negative stimuli in the WSAP task; WSAP_BER represents the endorsement rate for positive stimuli in the WSAP task; SST_TNR represents the ratio of completed negative sentences to total successfully unscrambled sentences in the SST; SRT_NT represents the scores for negative target sentences in the SRT; SRT_PT represents the scores for positive target sentences in the SRT; SRT_NT/SRT_PT represents the ratio of SRT_NT divided by SRT_PT.

4.3.3.2 SST

The ratio of completed negative sentences (*SST_TNR*) on the SST provided an index of negative interpretation bias (Rude et al., 2003; Holmes et al., 2009). Moderate positive correlations were found between *SST_NER* and depression at *T1* (r = .529, p = .024) and *T2* (r = .563, p = .015), suggesting the SST offers a moderately valid measure of negative bias associated with depression.

4.3.3.3 SRT

In this study, *SRT_PT* scores (range 0-60) was used as the primary counter-indicator of negative interpretation bias on the SRT, rather than *SRT_NT*, based on *SRT_PT* demonstrating greater reliability and validity as a measure of negative bias in previous sections. The SRT's oppositional design makes concurrent endorsement of both positive and negative targets unlikely, hence *SRT_PT* serves as a more direct indicator.

The *SRT_NT*/*SRT_PT* ratio was included as a secondary measure reflecting bias polarity, with higher ratios indicating more negative bias (Micco et al., 2013; Sfärlea et al., 2020). However, this ratio can be artificially inflated by low *SRT_PT* scores. To avoid distortion from indiscriminate responses, outliers were screened and removed beforehand, as described in previous sections.

Supporting *SRT_PT* 's validity as the counter-indicator, *SRT_PT* demonstrated significant negative correlations with depression severity at *T2* (r = -.591, p = .010), whereas *SRT_NT* did not. The *SRT_NT/SRT_PT* ratio also correlated with depression at *T2* (r = .495, p

= .037). Therefore, compared to *SRT_NT*, *SRT_PT* provides the more valid measure of negative interpretative bias among the SRT indices.

Replicated analyses without the outliers in their own indices yielded the same significance levels (Table 4.3).

Table 4. 3 Pearson Correlation Between Depression (BDI-II Scale) and Negative Bias Indicators (WSAP, SST, SRT) to Test Validity, Without Respective Outliers

Task	WSAP_NER	WSAP BER	SST_TNR	SRT_NT	SRT_PT	SRT_NT/ SRT_PT
Day 1 (Time 1)	r = .568 (p = .022)	r = -.012 (p = .965)	r = .525 (p = .031)	r = -.176 (p = .514)	r = -.018 (p = .947)	r = .007 (p = .981)
Day 14 (Time 2)	r = .775 (p < .001)	r = .551 (p = .027)	r = .567 (p = .018)	r = .264 (p = .324)	r = -.642 (p = .007)	r = .597 (p = .015)

4.3.4 Differences on Cognitive Assessment by Language

Shapiro-Wilk tests indicated most scale scores were normally distributed (ps > 0.05), meeting the assumption for independent samples *t*-tests.

Independent sample *t*-tests found no significant differences between language groups for depression severity, SST negative ratios, *SRT_NT*, or most WSAP ratios across timepoints (all ps > .05).

The only significant difference was on WSAP negative endorsement ratios (*WSAP_NER*), with the Mainland China group showing greater negative bias compared to Hong Kong at *T1* (p = .022) and marginally at *T2* (p = .050).

Table 4. 4 Paired Samples t-Test Across Language to Check Non-Difference in Assessment Evaluation

Variables	Language	Shapiro-Wilk			t-test Equality of Means			
		Statistic	df	Sig.	t	df	MD (Std. Error)	Sig.
DS_MS_1	Mandarin	.935	7	.596	1.707	15.978	4.909 (2.876)	.107
	Cantonese	.968	11	.869				
DS_MS_2	Mandarin	.808	7	.049	1.269	15.687	4.195 (3.305)	.223
	Cantonese	.861	11	.059				
SST_TNR_1	Mandarin	.940	7	.635	1.266	14.725	.131 (.104)	.225
	Cantonese	.962	11	.792				
SST_TNR_2	Mandarin	.896	7	.309	1.862	13.859	.192 (.103)	.084
	Cantonese	.971	11	.895				
SRT_PT_1	Mandarin	.875	7	.204	-1.007	15.996	-2.653 (2.635)	.329

	Cantonese	.947	11	.605				
SRT_PT_2	Mandarin	.823	7	.068	-1.722	15.005	-5.378 (3.124)	.106
	Cantonese	.906	11	.219				
SRT_NT_1	Mandarin	.986	7	.983	.131	11.879	.326 (2.495)	.898
	Cantonese	.984	11	.984				
SRT_NT_2	Mandarin	.845	7	.110	-.567	15.988	-1.498 (2.644)	.579
	Cantonese	.913	11	.266				
WSAP_NER_1	Mandarin	.893	7	.290	2.587	14.001	.158 (.061)	.022
	Cantonese	.782	11	.006				
WSAP_NER_2	Mandarin	.747	7	.012	2.133	14.968	.188 (.088)	.050
	Cantonese	.931	11	.424				
WSAP_PMR_1	Mandarin	.700	7	.004	.934	15.109	.045 (.049)	.365
	Cantonese	.952	11	.674				
WSAP_PMR_2	Mandarin	.930	7	.549	.271	10.784	.026 (.098)	.791
	Cantonese	.934	11	.456				
SRT_NBR_1	Mandarin	.870	7	.187	.949	13.653	.280 (.295)	.359
	Cantonese	.610	11	<.001				
SRT_NBR_2	Mandarin	.933	7	.574	1.193	13.696	.293 (.245)	.253
	Cantonese	.849	11	.041				

4.4 Summary and Discussion

In this study, Chinese versions of cognitive bias assessments including the Beck Depression Inventory-Second Edition (BDI-II), Word-Sentence Association Paradigm for Depression (WSAP-D), Scrambled Sentence Task (SST), and Similarity Rating Task (SRT), were translated and digitized in mobile application. To ensure data quality, missing values were handled, and outliers were screened. Reliability analyses using Cronbach's α and split-half reliability showed good internal consistency for BDI-II depression severity score (*DS_MS*), SST negative unscrambling ratio (*SST_TNR*), and WSAP-D negative endorsement rate (*WSAP_NER*). SRT positive target responses (*SRT_PT*) exhibited greater stability compared to other SRT indices.

The study aimed to test the following hypotheses:

H4.1: The cognitive bias measures and their indicators will demonstrate temporal stability.

H4.2: The negative bias indices from the SST (*SST_TNR*) and WSAP-D (*WSAP_NER*) are expected to correlate strongly with BDI-II scores at both timepoints.

H4.3: The SRT indices may not correlate strongly with BDI-II scores.

H4.4: Positive bias indicators (*SRT_PT*) in SRT may not negatively correlate with BDI-II scores.

Correlational analyses between depressive symptom severity and cognitive bias indices revealed that *WSAP_NER* and *SST_TNR* were significantly associated with *DS_MS*, supporting their validity in assessing depression-related negative interpretation biases (H4.2 supported). In contrast, SRT indices did not consistently correlate with depression severity (H4.3 supported). Furthermore, *SRT_PT* did not consistently exhibit significant negative correlations with *DS_MS* (H4.4 supported). Apart from *WSAP_NER*, independent *t*-tests showed no significant difference between language versions on most indices. Overall, the translated assessments demonstrated acceptable reliability (H4.1 partially supported) and validity.

The main limitation of this study is the small sample size. However, the findings are consistent with previous studies (Würtz et al., 2022; Gonsalves et al., 2019), suggesting the measures are reliable. In upcoming studies, the reliability and validity of these measures will be further reported. This chapter discussed the assessment tools for negative interpretation bias and depression severity used in the thesis and explained the coding methods, providing a foundation for Study 1 in Chapter 5 and Study 3 in Chapter 7.

Chapter 5: Effect of eiIBM_RobotV1 (Study 1)

This chapter explores the experiences caused by three types of robots (audio bot, telepresence robot and chatbot) delivering the traditional intervention of the elaborative interpretation bias modification(**eiIBM)** online. Specifically, the delivery mediums in this study were Audio, Video and Text. Through examining the experiential variables derived from I-PEFiC difference across time and thus effect on cognitive outcomes (BDI-II, WSAP-D, SST, SRT), this chapter empirically investigate influence of robot's type on therapy across time, contributing to the knowledge of incorporating the social robot into the standardized empirical-evidenced therapy.

5.1 Introduction

This study incorporated three types of robots (audio bot, telepresence robot, and chatbot) into an elaborative interpretation bias modification (**eiIBM**) program to understand users' experiences with different robotic mediums (*Audio*, *Video* and *Text*) delivering the intervention. Building on the methodology in Chapter 3, an initial version of the robot-delivered **eiIBM** (**eiIBM_RobotV1**) was developed to address three research questions:

RQ1.1: How do user perceptions differ between interacting with different types of robots (dissimilarity hypothesis) or are the perceptions largely similar (similarity hypothesis)?

RQ1.2: Given **eiIBM_RobotV1**, do the intervention outcomes differ between interactions with different robot types?

RQ1.3: How do user perceptions influence intervention outcomes?

The study aimed to understand the effect of robot-delivered intervention (**eiIBM_RobotV1**) on user experience and outcomes, as well as how the experience influences intervention outcomes.

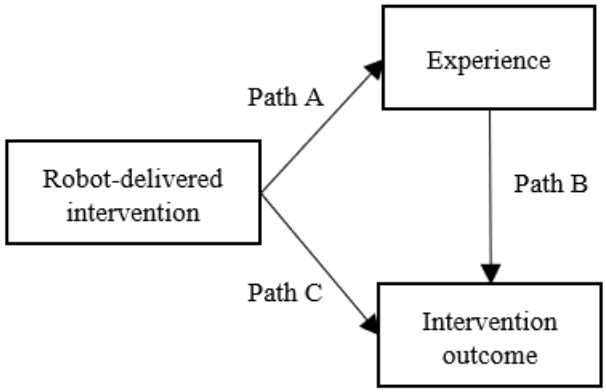

Figure 5. 1 Overview of Research Model for Study 2

Prior research indicates physical robots are often preferred for emotional support due to social cues that foster companionship and appeal (Li, 2015; Bainbridge, Hart, Kim, & Scassellati, 2010), and they effectively mitigate psychological distress (David & David, 2022; Law et al., 2022; Alemi et al., 2015; Pollak et al., 2022). Multimodal conversational agents also enhance social presence and personalization compared to text only, improving intervention outcomes through more human-like experiences (Li et al., 2023). This suggests multimodal robots should enable better user experiences and intervention effectiveness. The initial assumption was that robots with more modalities would contribute more to positive experience (Path A in Figure 5.1) and intervention outcomes (Path C in Figure 5.2).

However, recent studies (Bickmore et al., 2018; Vaidyam et al., 2019) show text-based chatbots can excel for certain goals like promoting healthy behaviors compared to voice-based chatbots. Huang et al. (2023) also found that the audio tutor outperformed a physical robot in reducing negative moods during mediation. This implies the effectiveness of different modalities may depend on context and goals, rather than assuming more modalities always lead to better performance. Robots with more modalities could potentially have no effect or even detrimental impacts on the intervention outcome (Path C in Figure 5.2).

Bickmore et al. (2018) proposed a text-only chatbot was more effective for promoting healthy eating than a voice-enabled bot because text kept users focused on the health messaging. In Huang et al. (2023), the physical robot distracted from meditation whereas an audio tutor reduced moods more. This suggests below-expectation outcomes from increased robot modalities could stem from either the unconscious perception (attracting more attention) or conscious perceptions (feeling focused) during the interaction, shaping experience. The experience may involve both affective and reflective aspects when interacting with the relational agent. Therefore, the effect of modalities on experience is not always positive (Path A in Figure 5.2).

Since experience can contribute positively or negatively to intervention outcomes, there may be a trade-off of effects on Path C. This promoted re-examining relationships among the robot-delivered intervention, experience, and intervention outcome in the context of robotic therapy. For Path A, Hoorn & Huang (2024) highlight task-contingency as a key to eliciting positive robot experience.

Task-contingency derived from Task-Technology Fit (TTF; Goodhue & Thompson, 1995) is reframed in Hoorn and Huang (2024) as that "Affordance should be Relevant to user goals" (see Chapter 2). With a good task-contingency, the relevance and valence would be higher (in either polarity) due to affordance from the combined feature set of tasks (robots' task) and technology (robot) compared to separately. Per I-PEFiC (van Vugt et al., 2009; Hoorn, 2015b), when users agree with relevance and valence, higher use intention and engagement are established.

The target audience was depressed young adults claiming emotional distress impacts their quality of life. Their common goals were to relieve distress and have a friendly interaction experience. Depressed individuals typically have low mental energy and motivation, so interactions requiring less mental attention may suit their needs. While they

could have other specific goals (e.g., dealing with relationship or financial triggers), in this therapeutic context, those goals had little ambition for participants to achieve after understanding the research background, which was disclosed before registration.

Other goals such as being comfortable or having thoughts challenged were individual differences in goal achievement strategies, but all relate to relieving distress. Accordingly, **eiIBM_RobotV1** was designed to address these common goals. Therefore, the designed affordance of **eiIBM_RobotV1** (including **eiIBM** implementation and **eiIBM** delivering agent with corresponding behavior) are "intervention delivery" and "ease-of-use interaction" corresponding to two main bodies, respectively.

In I-PEFiC, participants compare the affordance with goals to determine the relevance and valence evaluation. Each affordance is compared to goals sets (composed of one or more goals), contributing to the use intention and engagement with **eiIBM_RobotV1**.

The hypotheses about experience with the three **eiIBM_RobotV1** robots (RQ1.1) are (also see Figure 5.2):

H 5.1: Equal affordance of intervention delivery is perceived from audio bot, telepresence robot and chatbot.

H5.2: Equal relevance and valence to emotional distress reliver is perceived, given the equal affordance.

H5.3: Telepresence and audio bots (versus chatbot) are perceived higher in ease-of-use interaction affordance.

H5.4: Higher relevance and valence for ease-of-use affordance are perceived from telepresence robot and audio robot (versus chatbot), given the higher perceived ease-of-use intervention affordance.

H5.5: Telepresence and audio robot (versus chatbot) gain more use intention and engagement, given the equal relevance and valence for emotional distress reliver but higher for ease-of-use.

H5.6: Telepresence robot and audio robot (versus chatbot) gain more satisfaction from higher use intention.

No difference was proposed in ease-of-use affordance between telepresence and audio bots because robot appearance provides visual attraction without facilitating ease-of-use and can be ignored without impeding exercise.

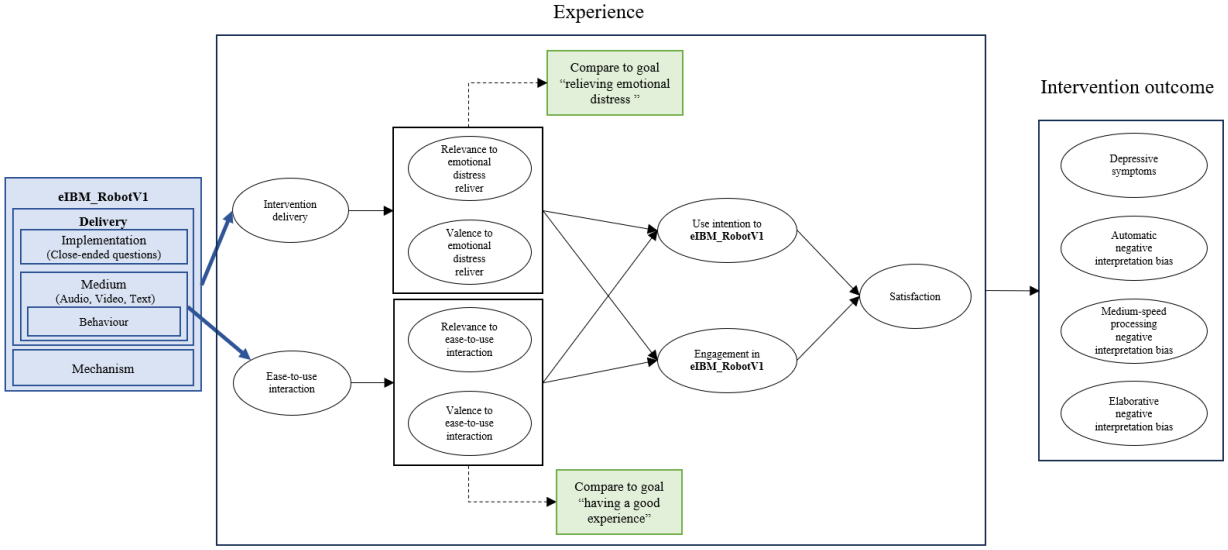

Figure 5. 2 The Hypothesis Overview for Study 2

For Path C, as the **eiIBM** mechanism was maintained in the robot version, the RQ1.2 hypothesis is:

H5.7: Telepresence and audio bots (versus chatbot) are more effective in reducing depression and negative interpretation bias.

For Path A to Path B, Chapter 2 describes that positive engagement is the effective **eiIBM** mechanism. Positive imagination, a specific exercise form, requires trust. Audio and

telepresence bots mobilize participants' multi-sensory channels, facilitating imagination engagement. Thus, the RQ1.3 hypothesis is:

H5.8: Participants with more positive use intention and engagement experience greater reductions in depression and negative bias.

To address the questions and test hypotheses, a between-subjects (**Medium**: *Video* versus *Audio* versus *Text*) repeated measures design was conducted.

5.2 Methods

5.2.1 Participant

This study targeted depressed young adults in Hong Kong who fulfilled the following inclusion criteria: 1) resident in Hong Kong, 2) aged 18-30; 3) able to read Chinese and understand spoken Cantonese or Mandarin; 4) access to Internet and WhatsApp; 5) showing at least four depressive symptoms lasting two weeks; 6) no drug addiction. Exclusion criteria were: 1) having arrangements disrupting usual routines during the study (e.g., traveling)

Participants were recruited from January to February 2023 via campus notice boards and social media (e.g., Facebook, Goop, Dcard). Interested individuals completed an online screening (Figure 5.3) through the QuestionPro link and those meeting criteria were contacted by researchers for confirmation.

Screening question 篩選問題

In the past two week, 在過去的兩周中

a. Have you felt low or difficult to be happy most of the day, almost every day?
   你是否幾乎每日大部分時間，都感到心情低落或難以開心？
b. Do you have little interest in anything, or less motivated to do anything most of the day?
   你是否幾乎每日大部分時間沒有動力做事？
c. How often have you experienced the following?
   你是否經常出現以下情況？ （可選多項）
   - Appetite changes, such as poor appetite or overeating
     食慾改變，如胃口變差或過度進食
   - Insomnia or poor sleep quality
     失眠或睡眠質量差
   - Talking or moving more slowly than usual, or fidgeting
     說話或行動變得比平日緩慢，或坐立不安
   - get tired easily
     容易疲累
   - difficulty concentrating
     難以集中精神
   - Decreased self-confidence, or blame yourself
     自信心下降，或責怪自己
   - Having thoughts of not wanting to live or suicidal
     有不想生存或自殺念頭
d. Does the above situation bother you significantly?
   上述情況是否對你做成明顯困擾？
e. Does the above situation have a significant negative impact on your life, such as studies, work, social interaction, etc.?
   上述情況對你的生活，例如學業、工作、社交等，是否做成明顯的負面影響？
f. Do you have drug dependence? If yes, which kind of drug?
   你是否有藥物依賴？如果有的話，是哪種藥物？

Figure 5. 3 Screening Questions for Participants

Eligible participants signed consent after reading an information sheet detailing the study aims, duration, involvement, randomization, and incentives. Finally, 49 depressed Hong Kong residents (M_{age}= 22.71, SD_{age}= 3.30, 38 Female, 35 Cantonese speaker) completed the experiments and received HK$350 ParkShop coupons as compensation. No participants exhibited severe acquiescence bias in after-experiment questionnaire (Jackson & Messick, 1965; Javeline, 1999).

5.2.2 Design

The study employed a between-subject (**Medium**: *Video* /*Audio*/ *Text*) repeated measures (**Time**: *T1*/ *T2* / *T3*) design with a pre- (*pretest*) and post-assessment (*posttest*). Eligible participants were randomly assigned to one of three **medium** conditions, controlling for age, gender, languages, and initial depression severity. The Video condition used video calls to emulate a telepresence robot; Audio used audio calls for an audio bot; Text used messaging for a chatbot.

The 2-week experiment involved two assessments and six robot-guided **eiIBM_RobotV1** sessions, with a minimum of 2 days between sessions. After the 1st, 3rd and 5th sessions, participants completed an experience questionnaire. Pre-post assessments were administered on **IBMTest@POLYUSD** mobile app, while **eiIBM_RobotV1** sessions was delivered via WhatsApp.

5.2.3 Procedure

After registration, researchers created **IBMTest@POLYUSD** accounts for eligible participants, personalizing language (Traditional/Simplified Chinese) and OS (iOS/Andriod) settings. The participation procedure is in Table 5.1.

In the preparation phase, researchers scheduled the 2-week experiment and sent instructions to install **IBMTest@POLYUSD** and complete pre-assessments the day before. Participants were to complete tasks sequentially without interruption, resting only between tasks. Inconsistent or unreasonable responses could lead to suspension. Researchers checked pre-assessment quality and scheduled the first session within 2 days. The remaining sessions were scheduled individually or together, with at least 2 days between each. Participants were informed of random assignment to *Text*, *Audio*, or *Video* robots, not needing camera/microphone on, and responding via text. An **eiIBM** mechanism video was also provided.

Reminders were sent the day before each session. For the first *Audio*/*Video* session, additional instructions covered earphones, minimizing video, turning off camera/microphone, and text responses.

Thirty minutes prior, participants were reminded to prepare a quiet environment and earphones (*Audio*/*Video*). Before the 1st, 3rd and 5th sessions, researchers noted a 10-minute

post-session questionnaire in case the participants left WhatsApp immediately after the call ended.

Sessions began with researcher confirming participant information, qualified environment, and full 30-minute engagement. They understood that early departure or unresponsiveness could suspend the experiment without compensation. No-shows were rescheduled within 4 days of the last session, or the project was suspended if unresponsive.

During sessions, robots guided participants through **eiIBM_RobotV1** (detailed in 5.2.4) exercise. After the 1st, 3rd and 5th session, participants were guided to fill out a questionnaire inquiring about the perception and experience of the interaction. Researchers announced completion after checking the questionnaire quality or directly if none that day.

Post-assessments were completed within 2 days of all sessions, following pre-assessment procedures. Researchers checked assessment quality, confirmed experiment completion, and scheduled reward pick-up. Coupons were delivered in-person or mailed per preference, with participants signing digital receipts.

Table 5. 1 Procedure of Participation in Study 1

Phase	Researchers' action	Participants' action
Before pre-assessment (1 day before scheduled 2-week period)	Sent links to install assessment app and start pre-assessments via WhatsApp. Instructed to complete tasks sequentially without stopping within a task but can rest between tasks. Reminded of right to suspend participation if unreasonable response times or inconsistencies found.	Received messages. Installed app. Completed pre-assessment tasks as instructed.
After pre-assessment	Checked assessment quality. Scheduled first exercise session within 2 days of pre-assessment. Confirmed first session timeslot. The remaining 5 timeslots scheduled session-by-session or all at once. Reminded of ≥2 days between sessions. Informed of random assignment to robot through message, call or video call. Sent video explaining principles.	Received confirmation of qualification. Scheduled first exercise session. Received info on exercise format. Watched explanation video.
Before session	Reminded of next day's scheduled session or rescheduled if needed. Sent instructions for audio/video groups.	Confirmed participation or rescheduled. Reviewed instructions.
30 mins before session	Reminded to prepare quiet environment. Audio/video groups were reminded to prepare earphones. Noted 10 min	Prepared for session.

	questionnaire after 1st, 3rd, 5th sessions. Checked environment.	
During exercise	Confirmed info and full 30 min engagement. Reminded leaving early or no response would suspend participation. Rescheduled or suspended project if no show. Conducted exercise session. After 1st, 3rd, 5th sessions, sent questionnaire. Checked responses. Announced completion.	Confirmed full engagement. Conducted exercise session. After 1st, 3rd, 5th sessions, sent questionnaire. Checked responses. Announced completion.
Before post-assessment	Assigned post-assessment tasks. Reminded to complete within 2 days.	Completed post-assessment.
After post-assessment	Checked validity. Scheduled coupon pick-up time.	Scheduled pick-up time.
End	Delivered coupons in-person or by mail. Had participant sign digital receipt.	Received coupons. Signed receipt.

5.2.4 Apparatus and Materials

5.2.4.1 Pre-and Post-Assessment Apparatus

IBMTest@POLYUSD, a translated and digitalized mobile app consisting of four cognitive tasks (BDI-II, SST, SRT, and WSAP-D), was used for pre- and post-assessments, as described in Chapter 4. Researchers managed participant accounts and monitored assessment progress using the **IBMTestManagement@POLYUSD** platform, also detailed in Chapter 4.

5.2.4.2 eiIBM Exercise Materials

5.2.4.2.1 Robot

The robot featured in the interaction videos was a NAO Humanoid Robot[1] model, renamed Zora (see Figure 5.4) developed by SoftBank Robotics. The two-syllable name Zora aligns with Chinese culture, making it more familiar and memorable. Standing 58 cm tall, Zora is equipped with sensors, controls, a voice synthesizer, an inertial board, and an Intel ATOM 1.6 GHz processor, with a battery life exceeding 1.5 hours. Zora's functionalities were manipulated using Choregraphe, SoftBank Robotics' software development kit, commonly

1 https://www.softbankrobotics.com/emea/en/nao

used in research and education, making it suitable for examining human-robot interactions. As NAO lacks a built-in Cantonese language library, the robot's audio was adapted into Cantonese and Mandarin using Google's text-to-speech service.

Figure 5. 4 NAO Robot Developed by Softbank Robotics

Figure 5.4 NAO Robot Developed by Softbank Robotics

5.2.4.2.2 Interpretation Bias Modification Task Stimuli

The task stimuli (scenarios) for **eiIBM_RobotV1** were adapted from the existing materials in Mathews and Mackintosh (2002), Dapprich et al. (2022) and Blackwell et al. (2022). These materials depict everyday situations that could potentially provoke negative emotions but are ultimately resolved positively or benignly. However, the original materials required adaptation. First, the scenarios from Blackwell et al. (2020) were in German, while those from Mathews and Mackintosh (2002) and Dapprich et al. (2022) were in English; all required translation to Written Mandarin and Cantonese. Second, Blackwell et al.'s (2022) scenarios were complete paragraphs with positive endings (e.g., "You must make a potentially difficult call and think about how it could happen when you enter the number. You feel greatly confident that he will run well."), rather than positively resolved scenarios with

missing words and corresponding comprehension questions (e.g., “You are asked to present about a topic without much preparation time. You present and people think you sound knowl_dgable. Did you sound knowledgeable?”).

To adapt the materials, 156 scenarios were translated into complete Cantonese paragraphs with a positive or benign resolution format. According to Beck (1967), negative interpretation sources consist of self-, future- and world-perspectives. Therefore, the selected scenarios could induce a negative mood from any of the perspectives. According to Beck (1967), negative interpretation sources consist of self-, future-, and world-perspectives. Based on this and inspection of the scenarios, self-perspective negativity was further divided into self-performance and self-feeling concerns; future-perspective negativity into hopelessness concerns on things and self; and world-perspective negativity into environmental and social concerns. A past-perspective including regret concerns was also added. These categories represent various sources of negative thinking and emotions.

Self-performance concerns represent the negative thinking about oneself, e.g., “I am a loser”; Self-feeling concerns represent getting into the negative mood, e.g., “I am upset.”; hopelessness concerns on things refer to the belief that things in the further will go worse. e.g., “things can only get worse!”; hopelessness concerns on self-focus on the belief that one’s own future is negative, e.g., “Nothing good for me can happen”; social concerns focus on the negative feeling induced by social interaction or social relationship. e.g., “no one values me”; environmental concern refers to the relationships to the objective stuff, e.g., “I hate the rain, it stops me from going out.”; regret concern refers mostly to the negative feeling when looking back on the experience/memory. Hopelessness concerns for self are different from performance concerns about self as the latter focuses on the uncertainty about the result of what I have done, the former is more about those results that have not yet happened.

Two research assistants categorized each scenario into one of the seven types and rated the level of induced negativity (1 = little, 2 = medium, 3 = high) and the difficulty of generating a positive completion (1 = easy, 2 = medium, 3 = difficult). They also created possible attributions for each scenario's positive completion. The categorizations and ratings were summarized, and appropriate attributions for positive completion were decided by the author. This process yielded 123 scenarios (21 for the former three sessions and 20 for the latter three sessions), divided into materials for six exercise sessions. Scenarios with lower negativity and easier positive completion were assigned to earlier sessions, while those with higher negativity and difficulty were assigned to the later sessions.

The scenarios were transformed into the format required for the **eiIBM** exercise, including a positively resolved ambiguous scenario with one word missing, a comprehension question, and an attribution to explain the positive resolution. Corresponding images with descriptive titles were added to enhance imagery (Holmes & Mathews, 2005; Holmes et al., 2008).

A sample adapted scenario is as follows:

[image with title] A image with the scenario title

[Scenario with one word missing] You walk on the street; you see your neighbor is not far away from you. You call him by name and try to greet him, but he does not answer you. You may be thinking it is because he did not ____ you. (你走在擁擠的大街上，你看見你的鄰居在離你不遠處。你叫他的名字試圖跟他打招呼，但是他沒有回應你。你想這可能是因為他沒有__到.)

[Comprehension question] Is your neighbor deliberately ignoring you? (你的鄰居是故意無視你嗎?)

[Attribution] The street was too noisy. He was busy rushing to a meeting, so he did not hear your voice or see you. (街上太嘈雜了，他正忙著趕去開會，所以並沒有聽到你的聲音也沒有看到你。)

5.2.4.2.3 Six-Session Intervention Program- eiIBM_RobotV1

The six-session **eiIBM_RobotV1** intervention program was designed with interactive content referring to existing Internet-based self-help IBM programs for depression and anxiety disorders (e.g., Titov et al., 2010), aligning with evidence-based traditional IBM mechanisms. The protocol of **eiIBM_RobotV1** session was shown in Figure 5.5. The exercise began with the researcher initiating contact through a video call, audio call, or text message, depending on the participant's group (*Video*, *Audio*, or *Text*). For *Video* groups, participants were instructed to turn off their camera and microphone after answering the call. The robot on-screen checked if participants could hear and see it, sending a repeat text message if the audio was unclear. Technical issues were addressed by the robot suggesting fixes or by researcher intervention. Once audio and video settings were confirmed, the robot led the exercise, presenting scenarios orally with key text messages for *Audio* and *Video* groups, and through full text instructions for Text groups.

In the first session, robots provided a detailed self-introduction, explaining the exercise frequency, length, response types, and the use of imagery, including an imagery exercise. Subsequent sessions had briefer introductions, with Sessions 2 and 3 offering replay options and Sessions 4-6 assuming familiarity with the rules. Each exercise presented scenario titles, images, and descriptions with one missing word, with reminder instructions being skipped in later sessions when participants were familiar with the format. An example of oral instruction is shown in Table 5.2.

Table 5. 2 Example of Robot Oral Instruction in Study 1

Robot sent message	**Robot oral instruction**
Run into a neighbor  	Scenario one /
You walk on the street; you see your neighbor is not far away from you. You call him by name and try to greet him, but he does not answer you. You may be thinking it is because he did not ____ you. /	You walk on the street; you see your neighbors not far away from you. You call him by name and try to greet him, but he does not answer you. You may be thinking… Please fill in the missing words in the horizontal line and send it by message to me. (optional)

Participants sent the missing characters to the robots, receiving congratulations of "Good answer (答啱啦)" speaking and "Correct (正確)" text for correct answer and encouragement to try again if incorrect. The robot gently suggested the correct answer when participants had difficulty or failed in the second trail, e.g., "The possible answer could be *hear*" for the scenario described in Table 5.2. Following each scenario, the robot asked a comprehension question to consolidate the positive resolution, with a choice between "1" or "2" or a simple yes/no question. One example of the yes/no question based on the above scenario was the latter "Are your neighbors deliberately ignoring you?" Congratulations were given for answers in line with the resolved scenario, while confirmation statements were provided if participants rejected the positive resolution. For example, on the "Run into a neighbor" scenario (see Table 5.2), based on a "no" response, the audio or telepresence robots said, "Good answer (答啱啦)" and texted "Correct (正確)". But the robots did not say a "yes" answer was wrong directly; instead, they gave a confirmation statement like "Your neighbor is not deliberately ignoring you." The robot moved to a new scenario after

completing each one until finishing all in this session. If paused too long, the robot reminded the participants about the task. Finally, the robot announced the session's completion and bid farewell to the participants.

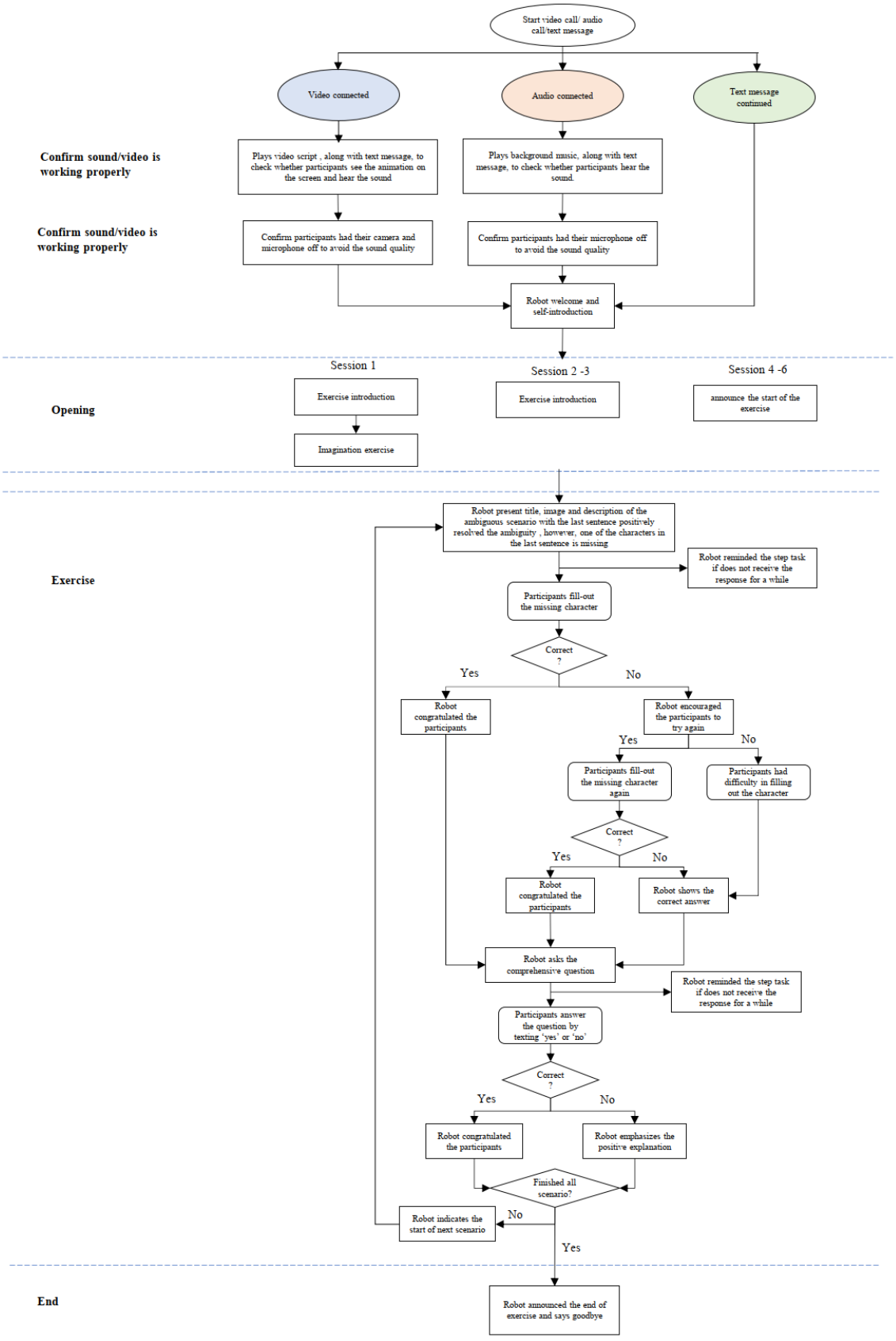

Figure 5.5 The Protocol of eiIBM_RobotV1

5.2.4.2.4 IBMOperationSystem@POLYUSD

The **eiIBM_RobotV1** exercise operation utilized a local website named **IBMOperationSystem@POLYUSD** to reduce variance in participant experience caused by operation issues. All text and audio scripts from audio and telepresence robots were pre-designed and pre-recorded, along with robot animations presenting the instructions.

The researcher prepared oral and text scripts for each item in Excel, and Python code read the scripts and transformed them into single audio files using the Microsoft Speech API-Azure. Compared to a mature male or female adult voice, or even a young girl's voice, the boy's voice was deemed more natural-sounding within the constraints of the available options. Therefore, parameters matching the Nao robot's appearance and natural-sounding feeling in Cantonese and Mandarin were set accordingly:
(Mandarin) Name: zh-CN-YunxiNeural; Pitch: +10%; Speed: 0.7f; Role: Boy; Style: cheerful
(Cantonese) Name: yue-CN-XiaoMinNeural; Pitch: 1%; Speed: =-12% Style: cheerful

Research assistants also recorded animations of the Nao robot aligned to each audio file, with diverse backgrounds for each session to increase reliability. This process resulted in audio and corresponding video files for each session, with audio files being merged into single session files having 5-second blanks between items. In total, 12 audio files and 12 video files were produced for the 6 sessions and 2 languages. A 1-minute idle animation of the breathing robot with flickering chest light was also recorded for waiting periods and embedded into the video files.

The **IBMOperationSystem@POLYUSD** website, created using JavaScript and HTML, was used to operate the audio and video clips within each session. The webpage consisted of a parent page and a child page on main and extended screens. The parent page

contained button links allowing researchers to control ready-to-play clips in the child page player, with corresponding text messages displayed at the bottom of the parent page for copying into WhatsApp. The environmental setup is shown in Figure 5.6.

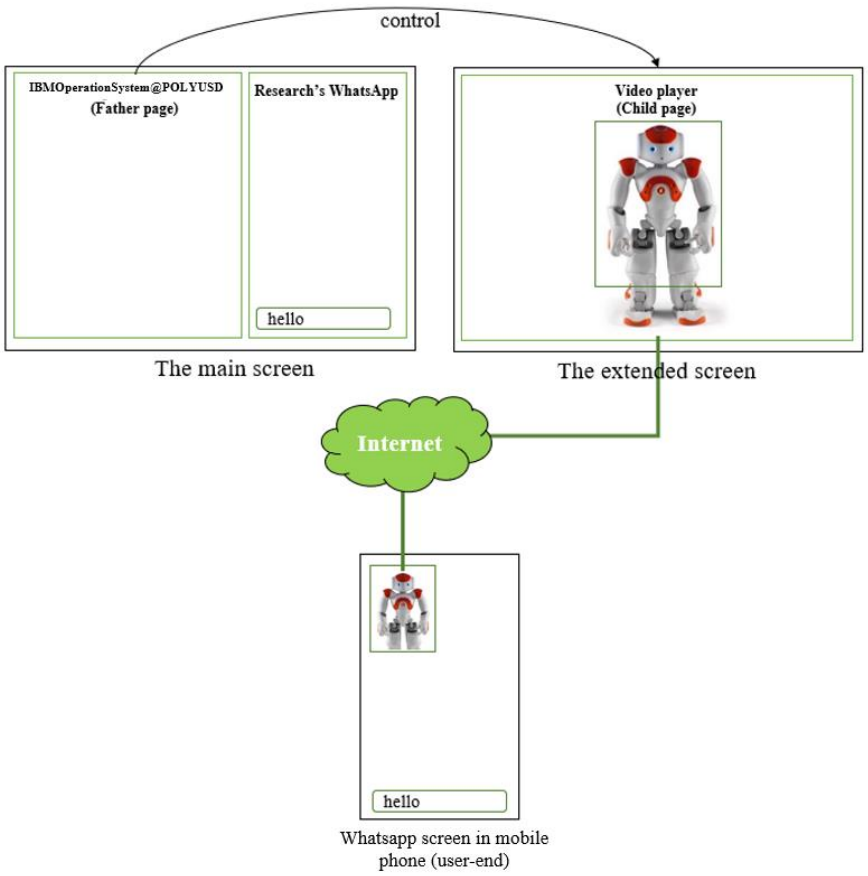

Figure 5.6 Environmental Setup for eiIBM_RobotV1

5.2.4.2.5 WhatsApp

WhatsApp was chosen as the communication platform between robots and participants due to its ability to meet the requirements of an ideal platform and its widespread use in Hong Kong. Zoom was considered but ultimately abandoned, as its phone interface does not allow simultaneous display of the video screen and chat area, potentially distracting participants from the exercise.

A pilot study was conducted with postgraduate students in the Multimedia & Entertainment Technology program taking Psychology of Design to test the setting's ability

to emulate live interaction. Each student experienced the exercise through text message, video call, and audio call, with none reporting suspicion that the robot responses were pre-recorded.

5.2.5 Measures

5.2.5.1 Cognitive Measurements

The same cognitive measurements from the preparation study were used: *DS_MS, SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*.

5.2.5.2 Experiential Measurements.

The experiential variables in the theoretical framework for Study 1 were measured at three points during the experience. The names of the variables and their relationships were derived from I-PEFiC model (Figure 5.7).

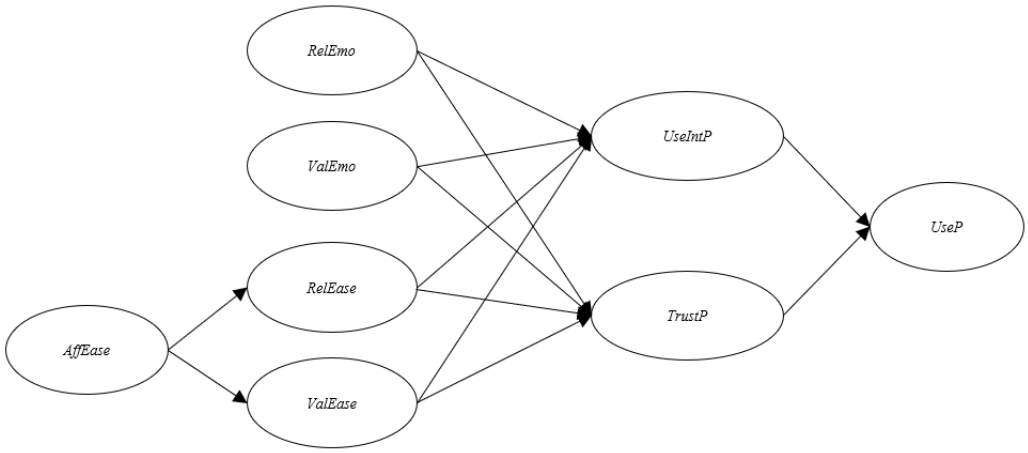

Figure 5.7 The Relationships of Experiential Variables Derived from I-PEFiC

It was expected that depressed participants would have the goals of relieving emotional distress (*GoEmo*) and having ease-of-use interaction (*GoEase*). Accordingly, **eiIBM_RobotV1** was designed with the affordance of "intervention delivery" (*AffEmo*) and "ease-of-use interaction" (*AffEase*). While *AffEmo* was explicitly perceivable through the

features of **eiIBM_Robot** across the three robots, *AffEase* differed by Medium and was therefore measured as a variable.

The comparison of *AffEmo* to *GoEmo* produced perceived relevance (*RelEmo*) and expectation of achieving the goal (*ValEmo*). In RelEmo, the concrete context of participant needs and *AffEmo* importance was considered fuzzy due to individual differences. Thus, the comparison was derived purely from the features of both eiIBM exercise format (*eiIBM_Implementation*) and robot (*eiIBM_Medium*).

AffEase 's relevance (*RelEase*) and valence (*ValEmo*) regarding *GoEase* explored facilitating influence of robots (*eiIBM_Medium*) on **eiIBM _RobotV1** program. Use intention to program (*UseIntP*), Engagement (*TrustP*), and satisfaction (*UseP*) were measured as holistic responses to **eiIBM_RobotV1**. Items for the experiential variables were generated from the literature sources in Table 5.3.

Table 5. 3 Items for Scales Measuring Experiential Variables in Study 1

Construct	Categories	Experience source	
Affordance	Affordance – Ease-of-use (*AffEase*)	Program (exercise format and robot)	*AffEase#_1*: Based on my experiences, the training program's interactions are clear and easy to understand *AffEase#_2*: Based on my experiences, I find that such training programs are easy to use *AffEase#_3*: Based on my experiences just now, I find that I can easily become proficient in the interaction process *AffEase#_4R*: Based on my experience just now, - it would take me a long time to get used to such a training program (R) *AffEase#_5*: Based on my experiences, I immediately understand how I should interact with the training program *AffEase#_6R*: Based on my experiences, the training program is a little difficult to use (R)
	Affordance – imagination (*AffImg*)	Exercise format	*AffImg#_1R*: During the training, Xiaozhi (guide) described a lot of scenes. For those scenes, it's hard for me to imagine those scenes happening to me *AffImg#_2*: During the training, Xiaozhi (guide) described a lot of scenes. For those scenes, - I can't relate myself to those scenes *AffImg#_3R*: During the training, Xiaozhi (guide) described many scenarios. For those scenes, - I can't relate myself to those scenes *AffImg#_4R*: During the training, Xiaozhi (guide) described many scenarios. For those scenes, - they're unfamiliar to me
Relevance			*RelEmo#_1*: I found this training program - it's what I need

	Relevance-program relevant to my goal of emotional reliver (*RelEmo*)	Program (exercise format and robot)	*RelEmo#_2*: I find such training programs – It's important to me *RelEmo#_3R*: I found that this training routine - it's not what I need *RelEmo#_4*: I found this training program - it matches my needs
	Relevance – robot relevant to the goal of completing the exercise (*RelEase*)	Robot	*RelEase#_1*: In terms of providing help, Xiaozhi (guide) - makes me complete tasks faster *RelEase#_2*: In terms of providing help, Xiaozhi (guide) - improves my efficiency *RelEase#_3*: In terms of providing help, Xiaozhi (guide) - helped me get involved in the training *RelEase#_4*: In terms of providing help, Xiaozhi (guide) - enhances the training effect *RelEase#_5R*: In terms of providing help, Xiaozhi (guide) - increased the difficulty of my training
Valence	Valence – expectation on the effect of emotion reliver (*ValEmo*)	Program (exercise format and robot)	*ValEmo#_1*: In terms of relieving emotional distress, I believe this training program - it makes me feel more comfortable *ValEmo#_2R*: In terms of relieving emotional distress, I believe this training program - it's not good for me *ValEmo#_3*: In terms of alleviating emotional distress, I believe that this training program - can change my situation *ValEmo#_4R*: When it comes to alleviating emotional distress, I'm confident that this training program will not make a difference
	Valence – expectation on the comfortable experience (*ValEase*)	Robot	*ValEase#_1*: The experience with Xiaozhi (guide), I felt - was interesting *ValEase#_2*: The experience with Xiaozhi (guide), I feel - is pleasant *ValEase#_3R*: The experience with Xiaozhi (the guide), I feel - is not happy *ValEase#_4*: In the experience with Xiaozhi (guide), I feel - it is a happy process
Use Intention	Use intention – program (*UseIntP*)	Program (exercise format and robot)	*UseIntP#_1*: Based on what I just experienced - I plan to continue using this training program in the future *UseIntP#_2*: Based on my experience just now, - in my daily life, I will try to use such training programs often *UseIntP#_3*: Based on what I have just experienced - I would recommend that other people in need try such training programs *UseIntP#_4R*: Based on what I just experienced - I don't plan to continue using such training programs
Trust	Engagement in the exercise (*TrustP*)	Program (exercise format and robot)	*TrustP#_1*: The following is a rating of trust in the training program. Overall, I - I think it's reliable *TrustP#_2*: The following is a rating of trust in the training program. On the whole, I - I can count on it *TrustP#_3R*: The following is a rating of trust in this training program. Overall, I - I'm not confident that it's going to work *TrustP#_4*: The following is a rating of trust in the training program. Overall, I - am confident that it can effectively change my negative interpretation of thinking
Satisfaction	Usefulness of the program (*UseP*)	Program (exercise format and robot)	*UseP#_1*: I think such a training program is useful *UseP#_2*: I think this kind of training program is valuable *UseP#_3R*: I think such training programs are - not valuable *UseP#_4R*: I think this training program is - useless

Note. # in variable name represent Time (1, 2, 3)

5.2.6 Reliability and Validity Analyses

Before conducting reliability analysis, counter-indicative items across three time points were reverse-coded (1→6, …, 6→1) for various scales, including two Affordance items (*AffEase#_4R, AffEase#_6R*), two Relevance items (*RelEmo#_3R, RelEase#_5R*), three Valence items (*ValEmo#_2R, ValEmo#_4R, ValEase#_3R*), one Use Intention item (*UseIntP#_4R*), one Trust item (*TrustP#_3R*) and two Usefulness items (*UseP#_3R, UseP#_4R*).

Cronbach's *α* were calculated to test scale reliability, followed by Principal Component Analysis (PCA) to assess construct validity. Reliability was established using observations from the first session, arguing that participants demonstrating consistent understanding on first exposure should show at least equal understanding subsequently.

5.2.6.1 Reliability and Validity Analyses with T1 Data

5.2.6.1.1 Reliability Analysis

Most scales in *T1* (Table 5.4) achieved very good reliability (Cronbach's α > .88) with all items included. They were *AffImg1, RelEase1, RelEmo1*, *ValEmo1*, *ValEase1*, *TrustP1*and *UseP. AffEase1, RealEase1* and *UseIntP1* reached good reliability (Cronbach's α > .85) after removing specific items (*AffEase1_4R* and *AffEase1_6R; RelEase1_5R; UseIntP1_4R*), while *UseIntT1* had acceptable reliability (Cronbach's α = .76) without *UseIntT1_2R* and could not be further improved by eliminating other items

Table 5. 4 Reliability of Scales for T1 in Study 1

Scale	Num of items	Items	Alpha / r	Standardized Alpha	Item mean	Variances of Item mean	*N*
AffEase1	4	*AffEase1_1, AffEase1_2, AffEase1_3, AffEase1_5 (AffEase1_4R and AffEase1_6R removed)*	*.900*	.900	5.05	.006	49

AffImg1	4	*AffImg1_1R*, *AffImg1_2*, *AffImg1_3R*, *AffImg1_4R*	.900	.900	4.40	0.008	49
RelEase1	4	*RelEase1_1, RelEase1_2, RelEase1_3, RelEase1_4* *(RelEase1_5R removed)*	.950	.950	4.60	0.005	49
RelEmo1	4	*RelEmo1_1*, *RelEmo1_2*, *RelEmo1_3R*, *RelEmo1_4*	.901	.905	4.25	0.002	49
ValEase1	4	*ValEase1_1, ValEase1_2, ValEase1_3R, ValEase1_4*	.889 (Improved to .902 if *ValEase1_3R* removed)	.900	4.56	0.007	49
ValEmo1	4	*ValEmo1_1, ValEmo1_2R, ValEmo1_3, ValEmo1_4R*	.903	.905	4.18	.054	49
UseIntP1	3	*UseIntP1_1, UseIntP1_2, UseIntP1_3* (*UseIntP1_4R* removed)	.856	.860	4.22	.032	49
TrustP1	4	*TrustP1_1, TrustP1_2, TrustP1_3R, TrustP1_4*	.884	.887	4.21	0.039	49
UseP1	4	*UseP1_1*, *UseP1_2*, *UseP1_3R*, *UseP1_4R*	.891	.894	4.59	0.004	49
UseIntT1	3	*UseIntT1_1, UseIntT1_3R, UseIntT1_4* *(UseIntT1_2R* Removed*)*	.765	.782	4.401	0.067	49

5.2.6.1.2 Validity Analysis

The PCA result with experiential items at *T1* and 8 fixed factors setting is shown in Table 5.5. The KMO (0.84) and significant Bartlett's test [$X^2(465) = 1532.23$, *Sig.* < .001] indicated a suitable sample size for factor analysis and appropriateness of the data for PCA (Tabachnick & Fidell, 2001; Field, 2014; Kaiser, 1974), expecting stable and interpretable factors. The PCA explained 84.64% of total variance. However, *RelEmo1* items spread over components, due to uncertainty in assessing 'need'. *RelEase1*, *UseIntP1*, *TrustP1* and *UseP1* items clustered into Components 1, 2, 3 and 6, respectively. To resolve cross-loading, *TrustP1_4* was removed, and items were retained as *UseP1_c* in Component 3, *TrustP1* in Component 8. Counter-indicated items of *UseP1* scale formed Component 6 and were assigned the name *UseP_ci*. *ValEase1* loaded onto Component 4, remaining *AffEase1* items onto Component 5, and remaining *ValEmo1* items onto Component 7. The second PCA with well-loading items showed a clear structure (Table 5.6)

Table 5. 5 Pattern Matrix with 8 Components on Experiential Variables for T1 in Study 1

Items	Component							
	1	2	3	4	5	6	7	8
AffEase1_1					**.692**			
AffEase1_2					**.862**			
AffEase1_3					**.831**			
AffEase1_5					**.808**			
RelEase1_1	**.918**							
RelEase1_2	**.868**							
RelEase1_3	**.741**							
RelEase1_4	**.778**							
RelEmo1_1							.507	
RelEmo1_2		.589						
RelEmo1_3R						.806		
RelEmo1_4							.585	
ValEase1_1				**.844**				
ValEase1_2				**.862**				
ValEase1_3R								
ValEase1_4				**1.077**				
ValEmo1_1							**.587**	
ValEmo1_2R							**.941**	
ValEmo1_3							**.603**	
ValEmo1_4R							**.558**	
UseIntP1_1		**.968**						
UseIntP1_2		**.997**						
UseIntP1_3		**.690**						
TrustP1_1								**.616**
TrustP1_2								**.611**
TrustP1_3R								
TrustP1_4			.651					
UseP1_1			**.896**					
UseP1_2			**.834**					
UseP1_3R						**1.006**		
UseP1_4R						**.934**		

Table 5. 6 Pattern Matrix with 8 Components on Experiential Variables for T1 in Study 1 (RelEmo Scale Removed)

Items	Component							
	1	2	3	4	5	6	7	8
AffEase1_1		**.688**						
AffEase1_2		**.900**						
AffEase1_3		**.869**						
AffEase1_5		**.805**						
RelEase1_1	**.971**							
RelEase1_2	**.974**							
RelEase1_3	**.775**							

RelEase1_4	**.872**									
ValEase1_1		**.827**								
ValEase1_2		**.820**								
ValEase1_4		**.911**								
ValEmo1_1				**.670**						
ValEmo1_2R				**.990**						
ValEmo1_3				**.687**						
ValEmo1_4R				**.552**						
UseIntP1_1			**.936**							
UseIntP1_2			**.932**							
UseIntP1_3			**.712**							
TrustP1									**.825**	
TrustP1_2									**.813**	
UseP1_1						**.811**				
UseP1_2						**.839**				
UseP1_3R							**.957**			
UseP1_4R							**.882**			

5.2.6.1.3 Reliability Analyses after PCA

Five experiential scales achieved good to very good reliability when run separately. *AffEase1*, *RelEase1*, *ValEase1*, and *ValEmo1* showed very good reliability (Cronbach's $\alpha > .88$), while *UseIntP1* demonstrated good reliability (Cronbach's $\alpha = .79$). For scales with only two items (*TrustP1*, *UseP1_i*, *UseP1_ci*), Spearman-Brown correlations showed strong and significant results, *TrustP1*: $r = .865$, $p < .001$, *UseP1_*i: $r = .822$, $p < .001$ and *UseP1_ci*: $r = .894$, $p < .001$ (Table 5.7). Results evident that participants had consistent understanding on the scales in *T1*.

Table 5. 7 Reliability Analyses of Scales for T1 After PCA in Study 1

Scale	Num of items	Items	Alpha / r	Standardized Alpha	Item mean	Variances of Item mean	*N*
AffEase1	4	*AffEase1_1, AffEase1_2, AffEase1_3, AffEase1_5*	.900	.900	5.05	.006	49
RelEase1	4	*RelEase1_1, RelEase1_2, RelEase1_3, RelEase1_4*	.950	.950	4.60	0.005	49
ValEase1	3	*ValEase1_1, ValEase1_2, ValEase1_4*	.902	.909	4.53	.007	49
ValEmo1	4	*ValEmo1_1, ValEmo1_2R, ValEmo1_3, ValEmo1_4R*	.903	.905	4.18	.054	49

UseIntP1	3	*UseIntP1_1, UseIntP1_2, UseIntP1_3*	.856	.860	4.22	.032	49
TrustP1	2	*TrustP1_1, TrustP1_2*	*r* = .865 (*p* <. 001)				49
UseP1_i	2	*UseP1_1*, *UseP1_2*	*r* = .822 (*p* <. 001)				49
UseP1_ci	2	*UseP1_3R*, *UseP1_4R*	*r* = .894 (*p* <. 001)				49

Given the I-PEFiC model posits each encoding is influenced by the previous response and evaluation, separate validity analyses (PCA) were conducted on items with high within-scale reliability at each timepoint (*T1*, *T2*, *T3*) to observe changes in factor validity throughout the experiment.

5.2.6.2 Reliability and Validity Analyses with T2 Data and T3 Data

Items remaining after *T1* analysis served as a benchmark for comparison with corresponding items from *T2* and *T3*. Table 5.8 summarizes reliability analyses results and component loadings in PCAs for each scale at *T1*, *T2* and *T3* independently. For *T2* and *T3*, scales achieving good to very good reliability included items corresponding to *T1*, except for *ValEmo2_2R*, which when removed improved *ValEmo2* Cronbach's α to .869.

Table 5. 8 Separate Reliability Analyses of Scales for T1, T2, and T3 in Study 1

Scale	***T1***		***T2***		***T3***		***N***
	Remaining items (component)	**Alpha / r**	**Remaining items (component)**	**Alpha / r**	**Remaining items (component)**	**Alpha / r**	
AffEase	*AffEase1_1*, *AffEase1_2*, *AffEase1_3*, *AffEase1_5* (Component 5)	.900	*AffEase2_1*, *AffEase2_2*, *AffEase2_3*, *AffEase2_5*, *AffEase3_4R*, *AffEase3_6R* (Component 2)	.922	*AffEase3_1*, *AffEase3_2*, *AffEase3_3*, *AffEase3_5*, *AffEase3_4R*, *AffEase3_6R* (Component 2)	.971	49
AffImg	*AffImg1_1R*, *AffImg1_2*, *AffImg1_3R*, *AffImg1_4R*	.900	*AffImg2_1R*, *AffImg2_2*, *AffImg2_3R*, *AffImg2_4R*	.854	*AffImg3_1R*, *AffImg3_2*, *AffImg3_3R*, *AffImg3_4R*	.928	49
RelEase	*RelEase1_1*, *RelEase1_2*, *RelEase1_3*, *RelEase1_4* (Component 1)	.950	*RelEase2_1*, *RelEase2_2* *RelEase2_3*, *RelEase2_4* (Component 4)	.914	*RelEase3_1*, *RelEase3_2*, *RelEase3_3*, *RelEase3_4* (Component 3)	.967	49

RelEmo	*RelEmo1_1*, *RelEmo1_2*, *RelEmo1_3R*, *RelEmo1_4*	.901	*RelEmo2_1*, *RelEmo2_2*, *RelEmo2_3R*, *RelEmo2_4* (Component 3)	.908	*RelEmo3_1*, *RelEmo3_2*, *RelEmo3_3R*, *RelEmo3_4* (Component 1)	.936	49
ValEase	*ValEase1_1*, *ValEase1_2*, *ValEase1_4* (Component 4)	. 902	*ValEase2_1*, *ValEase2_2*, *ValEase2_4* (Component 5)	.930	*ValEase3_1*, *ValEase3_2*, *ValEase3_3R*, *ValEase3_4* (Component 4)	.954	49
ValEmo	*ValEmo1_1*, *ValEmo1_2R*, *ValEmo1_3*, *ValEmo1_4R* (Component 7)	.903	*ValEmo2_1*, *ValEmo2_2R*, *ValEmo2_3*, *ValEmo2_4R* (Component 1)	.843->.869 if removed *ValEmo2_2R*	*ValEmo3_1*, *ValEmo3_2R*, *ValEmo3_3*, *ValEmo3_4R*,	.864	49
UseIntP	*UseIntP1_1*, *UseIntP1_2*, *UseIntP1_3* (Component 2)	.856	*UseIntP2_1*, *UseIntP2_2*, *UseIntP2_3*, *UseIntP2_4R* (Component 1)	.930	*UseIntP3_1*, *UseIntP3_2*, *UseIntP3_3* (Component 1)	.953	49
TrustP	*TrustP1_1*, *TrustP1_2*, *TrustP1_3R*, *TrustP1_4* (Component 8)	.884	*TrustP2_1*, *TrustP2_2*, *TrustP2_3R*, *TrustP2_4* (Component 6)	.923	*TrustP3_1*, *TrustP3_2* (Component 8)	r = .927 (p <. 001)	49
UseP	*UseP1_1*, *UseP1_2* (Component 3) *UseP1_3R*, *UseP1_4R* (Component 6)	.891	*UseP2_1*, *UseP2_2*, *UseP2_3R*, *UseP2_4R* (Component 3)	.946	*UseP3_1*, *UseP3_2* (Component 7) *UseP3_3R*, *UseP3_4R* (Component 1)	.951	49

Note. Loading components of *T1*, *T2* and *T3* were the results from Table 5.6, Table 5.9, Table 5.10.

The PCAs with the remaining *T2* and *T3* items in Table 5.8 and free fit settings only extracted three components, with many scales in the same components. Therefore, the fixed number of extracted components was set to eight and analyses were rerun. Results are shown in Tables 5.9 and 5.10

Table 5. 9 Pattern Matrix with 8 Components on Experiential Variables for T2 in Study 1

Items	Component							
	1	2	3	4	5	6	7	8
AffEase2_1		**.586**						
AffEase2_2		**.872**						
AffEase2_3		**.756**						
AffEase2_4R		**.972**						
AffEase2_5		**.733**						
AffEase2_6R		**.914**						
RelEase2_1				**1.000**				
RelEase2_2				**1.106**				
RelEase2_3								
RelEase2_4	.562							
RelEmo2_1			**.918**					
RelEmo2_2			**.668**					.698
RelEmo2_3R			**1.074**					

RelEmo2_4								
ValEase2_1					**.610**			
ValEase2_2					**.657**			
ValEase2_4					**.917**			
ValEmo2_1	**.638**							
ValEmo2_2R							.887	
ValEmo2_3	**.889**							
ValEmo2_4R								
UseIntP2_1	**1.012**							
UseIntP2_2	**.921**							
UseIntP2_3	**.644**							
UseIntP2_4R	**.556**							
TrustP2_1						**.768**		
TrustP2_2						**.678**		
TrustP2_3R								
TrustP2_4								
UseP2_1			**.636**					
UseP2_2			**.803**					
UseP2_3R								
UseP2_4R			**.577**					

Table 5. 10 Pattern Matrix with 8 Components on Experiential Variables for T3 in Study 1

Items	Component							
	1	2	3	4	5	6	7	8
AffEase3_1		**.792**						
AffEase3_2		**.761**						
AffEase3_3		**.888**						
AffEase3_4R		**.733**						
AffEase3_5		**.801**						
AffEase3_6R		**.880**						
RelEase3_1			**.937**					
RelEase3_2			**.996**					
RelEase3_3			**.830**					
RelEase3_4			**.749**					
RelEmo3_1	**.837**							
RelEmo3_2	**.755**							
RelEmo3_3R	**.708**							
RelEmo3_4	**.874**							
ValEase3_1				**.710**				
ValEase3_2				**.974**				
ValEase3_3R				**.527**				
ValEase3_4				**.828**				
ValEmo3_1								
ValEmo3_2R					.843			
ValEmo3_3	.798							
ValEmo3_4R						.720		
UseIntP3_1	**.902**							
UseIntP3_2	**.890**							

UseIntP3_3	**.789**			
TrustP3_1	.677			**.556**
TrustP3_2				**.619**
UseP3_1			**.584**	
UseP3_2	.578		**.609**	
UseP3_3R	**.576**			
UseP3_4R	**.644**			

The PCA results across *T1*, *T2*, and *T3* suggest a progressive understanding in participants' self-reported needs and preferences, which may have influenced responses and factor loading results. A consistent pattern was observed for *AffEase*, *ValEase* and *UseIntP* scales, with increasing Cronbach's α coefficients over time supporting an "understanding line" notion.

For *TrustP*, only *TrustP#_1* and *TrustP#_2* consistently clustered within the same component, highlighting construct stability with indicative items over time. The *ValEmo* scale showed an interesting trajectory, with items forming a distinct component at *T1*, but spreading across components by *T3*, suggesting changing perceptions on the valence of **eiIBM_RobotV1** relieving emotional distress.

Conversely, *RelEmo* items began forming a distinct component and highly correlating with *UseIntP* scale over time, implying participants gained clarity on their needs after repeated interactions. For *UseP*, differentiation between agreement and rejection of usefulness was initially less pronounced but re-emerged by *T3*, with rejection aligning with *UseIntP3* and *RelEmo.* It might be because participants had the central tendency to avoid choosing extremes (Paulhus, 1991) on the negative interpretation when they were not familiar with the interaction but became confirmed over time.

PCAs at different timepoints provide insights into the evolving understanding and correlation between experiential variables over time. Some constructs demonstrated stability, others required time to form confirmed evaluations, and some became less explicable and spread over components. Interrelatedness of variables also intensified.

5.2.6.3 Reliability and Validity Analyses with T2 items and T3 Items Corresponding to T1 Remining Items.

To make the experience comparable across time, *T2* and *T3* scale analyses and PCAs were conducted only on items corresponding to *T1* retained items. Table 5.11 shows all scales demonstrated good to very good reliability. Items in *T2* and *T3* corresponding to *T1* but absent in its sperate analysis were highlighted in grey and included.

Notable in *T2* was the clustering of *ValEase2* and *TrustP2* (Table 5.12), suggesting high correlation. Similarly, *ValEmo2* and *UseIntP2* grouped in Component 1, indicating substantial correlation. Although *RelEase2_3*, *RelEase2_4*, *ValEmo2_3* and *ValEmo2_4R* did not exhibit high PCA loadings, they were included for comparison. *UseP2_3R* and *UseP2_4R* did not form a distinct component but were maintained as a scale for comparison, with caution in interpretating analyses including these two items.

In *T3* (Table 5.13), *ValEmo3* items dispersed in the PCA but were used for comparison, noting weaker interpretive power. *UseIntP3*, *UseP3_i* and *UseP3_ci* fell under the primary component, underscoring a strong correlation between usefulness evaluations and willingness to use.

Table 5. 11 Reliability Analyses of Scales for T1, T2, and T3 with T1 as Benchmark in Study 1

Scale	*T1*		*T2*		*T3*		*N*
	Remaining items (component)	**Alpha / r**	**Remaining items (component)**	**Alpha / r**	**Remaining items (component)**	**Alpha / r**	
AffEase	*AffEase1_1, AffEase1_2, AffEase1_3, AffEase1_5* (Component 2)	.900	*AffEase2_1, AffEase2_2, AffEase2_3, AffEase2_5* (Component 2)	.918	*AffEase3_1, AffEase3_2, AffEase3_3, AffEase3_5* (Component 2)	.965	49
RelEase	*RelEase1_1, RelEase1_2, RelEase1_3, RelEase1_4* (Component 1)	.950	*RelEase2_1, RelEase2_2, RelEase2_3, RelEase2_4* (Component 5)	.914	*RelEase3_1, RelEase3_2, RelEase3_3, RelEase3_4* (Component 3)	.967	49

ValEase	*ValEase1_1, ValEase1_2, ValEase1_4* (Component 3)	.902	*ValEase2_1, ValEase2_2, ValEase2_4* (Component 3)	.930	*ValEase3_1, ValEase3_2, ValEase3_4* (Component 4)	.949	49
ValEmo	*ValEmo1_1, ValEmo1_2R, ValEmo1_3, ValEmo1_4R* (Component 5)	.903	*ValEmo2_1, ValEmo2_2R, ValEmo2_3, ValEmo2_4R* (Component 1)	.843	*ValEmo3_1, ValEmo3_2R, ValEmo3_3, ValEmo3_4R*	.864	49
UseIntP	*UseIntP1_1, UseIntP1_2, UseIntP1_3* (Component 4)	.856	*UseIntP2_1, UseIntP2_2, UseIntP2_3* (Component 1)	.902	*UseIntP3_1, UseIntP3_2, UseIntP3_3* (Component 1)	.953	49
TrustP	*TrustP1_1, TrustP1_2* (Component 8)	*r* = .865 (*p* <. 001)	*TrustP2_1, TrustP2_2* (Component 3)	*r* = .935 (*p* <. 001)	*TrustP3_1, TrustP3_2* (Component 6)	*r* = .927 (*p* <. 001)	49
UseP_i	*UseP1_1*, *UseP1_2* (Component 6)	*r* = .822 (*p* <. 001)	*UseP2_1*, *UseP2_2* (Component 4)	*r* = .902 (*p* <. 001)	*UseP3_1*, *UseP3_2* (Component 1)	*r* = .935 (*p* < .001)	49
UseP_ci	*UseP1_3R*, *UseP1_4R* (Component 7)	*r* = .894 (*p* <. 001)	*UseP2_3R*, *UseP2_4R*	*r* = .866 (*p* <. 001)	*UseP3_3R*, *UseP3_4R* (Component 1)	*r* = .907 (*p* < .001)	49

Table 5. 12 Pattern Matrix with 7 Components on Experiential Variables for T2 Corresponding to T1 in Study 1

Items	Component						
	1	2	3	4	5	6	7
AffEase2_1		**.669**					
AffEase2_2		**.935**					
AffEase2_3		**.875**					
AffEase2_5		**.859**					
RelEase2_1					.926		
RelEase2_2					.920		
RelEase2_3							.840
RelEase2_4							.771
ValEase2_1			**.640**				
ValEase2_2			**.743**				
ValEase2_4			**.985**				
ValEmo2_1	.565						
ValEmo2_2R						1.083	
ValEmo2_3	.894						
ValEmo2_4R							
UseIntP2_1	**.953**						
UseIntP2_2	**.739**						
UseIntP2_3	**.584**						
TrustP2_1			**.631**				
TrustP2_2			**.632**				
UseP2_1				**.935**			
UseP2_2				**.843**			
UseP2_3R							
UseP2_4R				**.626**			

Table 5. 13 Pattern Matrix with 7 Components on Experiential Variables for T3 Corresponding to T1 in Study 1

Items	Component						
	1	2	3	4	5	6	7
AffEase3_1		**.885**					
AffEase3_2		**.920**					
AffEase3_3		**.903**					
AffEase3_5		**.895**					
RelEase3_1			**.932**				
RelEase3_2			**.992**				
RelEase3_3			**.854**				
RelEase3_4			**.776**				
ValEase3_1				**.657**			
ValEase3_2				**.955**			
ValEase3_4				**.844**			
ValEmo3_1							
ValEmo3_2R							.651
ValEmo3_3	.929						
ValEmo3_4R					.980		
UseIntP3_1	**.937**						
UseIntP3_2	**1.081**						
UseIntP3_3	**.928**						
TrustP3_1	.523					**.606**	
TrustP3_2						**.659**	
UseP3_1	**.586**						
UseP3_2	**.737**						
UseP3_3R	**.567**						
UseP3_4R	**.590**						

5.2.6.4 Outlier Exploration

5.2.6.4.1 Exploration of Experiential Outliers

To identify potential outliers, means were calculated across items for the 8 remaining scales at each of the 3 timepoints. The resulting 24 mean values were prefixed with *M_* to distinguish from single-item values. Outlier analyses were then performed on each scale via boxplots with **Medium** factors. Boxplots indicated 10 potential outliers across all 8 variables and 3 timepoints in the 3 Medium conditions (Table 5.14). No extreme outliers were found. Table 5.14 shows the outlier distribution, with lower-end outliers marked with brackets.

As outliers are legitimate observations naturally occurring in populations and no hypotheses existed to exclude them, analyses were performed both including and excluding outliers to account for their potential to distort analyses and violate model assumptions. This resulted in two datasets at each timepoint: *T1* with outliers (*N1* = 49) and without (*n1* = 47), *T2* with outliers (*N2* = 49) and without (*n2* = 43), and *T3* with outliers (*N3* = 49) and without (*n3* = 44).

Table 5. 14 Outlier Distribution for Experiential Variables Across Medium and Time in Study 1

Condition	Audio			Video			Text		
Session	1	2	3	1	2	3	1	2	3
M_AffEase	/	/	/	/	/	/	(C_35)	/	/
M_RelEase	/	/	/	/	(C_27, C_30)	/	(C_35)	(C_35)	/
M_ValEase	/	/	/	/	/	/	/	/	/
M_ValEmo	/	/	/	/	(C_30)	/	/	/	/
M_UseIntP	/	C_15, C_9	(C_3)	/	/	/	/	/	/
M_TrustP	/	/	/	/	/	/	/	/	(C_35)
M_UseP_i	/	/	/	/	/	(C_26)	/	/	(C_35)
M_UseP_ci	/	/	/	(C_18)	(C_18)	C_20, (C_28)	/	/	/

***** Outliers in the lower end of the box are marked with a bracket

5.2.6.4.2 Exploration of Assessment Outcome Outliers

To ensure accuracy in intervention outcome analysis, outliers were checked for in each assessment task, categorizing non-serious and low effect participants per Chapter 4. Participants from Chapter 4 were included as the *Control* group. Table 5.15 flags the outliers from different assessments

For the BII-D2, no outliers emerged in the intervention groups. Aligning with Chapter 4, only C_18 had an extreme depressive symptom improvement value, resulting in one BII-D2 outlier in Study 1.

In the SST, two non-serious participants were identified for both pre-SST (E1_8, E1_11) and post SST (E1_11, E1_17) in the *Audio* group. E1_26 from the *Video* group and C_13 from the *Control* group were also non-serious in SST.

No outliers due to polarized response bias existed in the WSAP, but four potential non-serious outliers were found. E1_26 showed agreement then disagreement on negatives/positives, suggestive of random Set 2 responding. E1_35 showed balanced Set 1 responses but disagreed with all Set 2 negatives/positives, indicating biased 2 responses. C_11 and C_4 remained outliers from Chapter 4.

SRT Outliers

For SRT, #E1_37, #E1_45, #E1_12, #E1_15, #E1_18, #C_4 and #C_16 were outliers based on response patterns with bias 2 and 4. However, it was acknowledged that looking at overall rather than item-level ratios may have missed nuanced outliers that persisted as dataset noise. The distribution of outliers across **Medium** is shown in Table 5.16.

Table 5. 15 Outlier Distribution for SST, WSAP, and SRT in Study 1

Participant index	Pre-test				Post-test			
	Non-serious participants in SST [1]	Non-serious participants in WSAP [2]	Non-serious participants in SRT [3]	Low effect participants in SRT [4]	Non-serious participants in SST	Non-serious participants in WSAP	Non-serious participants in SRT	Low effect participants in SRT
E1_1	/	/	/	/	/	/	/	bias 1
E1_2	/	/	7 incorrect	/	/	/	/	/
E1_3	/	/	/	bias 1	/	/	/	/
E1_4	/	/	7 incorrect	bias 1	/	/	7 incorrect	bias 1
E1_8	7 incorrect	/	/	/	/	/	/	bias 1
E1_11	8 incorrect	/	/	/	7 incorrect		7 incorrect	bias 1
E1_12	/	/	7 incorrect 1 null	bias 4	/	/	5 incorrect 1 null	bias 1
E1_13	/	/	/	/	/	/		bias 1
E1_14	/	/	/	/	/	/		bias 1
E1_15	/	/	/	bias 4	/	/	/	/
E1_17	7 incorrect	/	/	/	/	/	/	bias 1
E1_18	/	/	9 incorrect	/	/	/	/	bias 4
E1_19	/	/	/	/	/	/	5 incorrect 1 null	bias 1
E1_24	/	/	/	bias 1	/	/	/	bias 1
E1_25	/	/	/	bias 1	/	/	5 incorrect 1 null	bias 1
E1_26	10 incorrect	NM_D:9 PE_D:8	5 incorrect 1 null	bias 1	/	/	5 incorrect 1 null	
E1_27	/	/	7 incorrect	/	/	/	/	bias 1
E1_28	/	/	/	/	/	/	/	bias 1
E1_30	/	/	/	/	/	/	/	bias 1
E1_31	/	/	6 incorrect	/	/	/	/	/
E1_33	/	/	6 incorrect	/	/	/	/	/
E1_34	/	/	/	/	/	/	/	bias 1
E1_35	/	PE_D:7	6 incorrect 1 null	/	/	/	/	/
E1_36	/	/	/	/	/	/	/	bias 1
E1_37	/	/	6 incorrect	bias 2	/	/	7 incorrect	bias 1
E1_38	/	/	/	bias 1	/	/	/	bias 1
E1_45	/	/	6 incorrect 1 null	bias 2	/	/	7 incorrect	/
E1_ 46	/	/	/	/	/	/	4 incorrect 1 null	/
E1_47	/	/	/	/	/	/		bias 1
E1_48	/	/	/	/	/	/	6 incorrect	bias 1
E1_49	/	/	/	bias 1	/	/	/	bias 1
C_4	/	/	/	/	/	PE_D: 10	/	bias 2
C_7	/	/	/	/	/	/	/	
C_11	/	PE_D: 8	/	/	/	/	/	/
C_12	/			bias 1				
C_13	/	/	6 incorrect	/	/	/	/	/
C_14	/	/	/	bias 1	/	/	/	/
C_16	/	/	6 incorrect	bias 2	/	/	4 incorrect 3 null	/

Note. NM_D means the rejection difference on negative words in two sets materials; PE_D means the endorsement difference on positive words in two sets materials; bias 1 means acceptance on positive statements without considering the context; bias 2 means unbiased rejection without considering the context; bias 3 means unbiased acceptance without considering context; bias 4 means acceptance on negative statements without considering context.

Table 5. 16 Outlier Distribution Across Medium for SST, WSAP, and SRT in Study 1

Items	Medium	Sample Size (total)	Outliers Index (Pre)	Outliers Index (post)	Sample Size (without outliers)
DS_MS	Audio	17	/	/	17
	Video	16	/	/	16
	Text	16	/	/	16
	Control	18	C_18	/	17
SST	Audio	17	E1_8, E1_11, E1_17	E1_11	14
	Video	16	E1_26	/	15
	Text	16	/	/	16
	Control	18	C_13	/	18
SRT	Audio	17	E12, E1_15	/	15
	Video	16	E1_18	E1_18	15
	Text	16	E1_37, E1_45	/	14
	Control	18	C_16	C_4	17
WSAP	Audio	17	/	/	17
	Video	16	E1_26	/	15
	Text	16	E1_35	/	15
	Control	18	C_11	C_4	16

5.3 Results

5.3.1 Demographics

To assess the success of random assignment, Chi-square tests were considered. As one assumption is that all variables are categorical, the continuous *Age* variable was recoded into categorical *Age range*, and the cognitive variable *Depression* was recoded into categorical *Depression Level*. This resulted in demographic distributions for the dataset with outliers included (Table 5.17) and datasets with outliers excluded (Table 5.18) at *T1*, *T2*, and *T3*. However, over 20% of cells had expected counts less than five (e.g., *Male* cells over Medium), violating the assumption for Pearson Chi-square test.

Table 5. 17 Demographic Distribution Over Medium with Outliers in Study 1

Demographics	Audio (n = 17)	Video (n = 16)	Text (n = 16)	Overall (n = 49)
Gender				
Female	13 (13.2*)	12 (12.4*)	13 (12.4*)	38
Male	4 (3.8*)	4 (3.6*)	3 (3.6*)	11
Depression Level				
0-14: Minimal	0 (0*)	0 (0*)	0 (0*)	0
14-20: Mild	3 (2.4*)	2 (2.3*)	2 (2.3*)	7
20-29: Moderate	6 (5.6*)	5 (5.2*)	5 (5.2*)	16
29-63: Severe	8 (9.0*)	9 (8.5*)	9 (8.5*)	26
Age (years)				
20 and below	7 (5.2*)	5 (4.9*)	3 (4.9*)	15
21-24	7 (8.3*)	7 (7.8*)	10 (7.8*)	24
25-29	3 (3.1*)	3 (2.9*)	3 (2.9*)	9
30 and above	0 (.3*)	1 (.3*)	0 (.3*)	1
Language				
Mandarin	6 (4.9*)	4 (4.6*)	4 (4.6*)	14
Cantonese	11 (12.1*)	12 (11.4*)	12 (11.4*)	35

Note. Expected frequencies for each cell.

Table 5. 18 Demographic Distribution Over Medium Without Outliers at Different Timepoints in Study 1

Time points	***T1***			***T2***			***T3***		
Medium	Audio (n = 17)	Video (n = 15)	Text (n = 15)	Audio (n = 15)	Video (n = 13)	Text (n = 15)	Audio (n = 16)	Video (n = 13)	Text (n = 15)
Gender									
Female	13 (13.0)	11 (11.5)	12 (11.5)	11 (11.2)	9 (9.7)	12 (11.2)	12 (12.4)	10 (10.0)	12 (11.6)
Male	4 (4.0)	4 (3.5)	3 (3.5)	4 (3.8)	4 (3.3)	3 (3.8)	4 (3.6)	3 (3.0)	3 (3.4)
Depression Level									
0-14: Minimal	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
14-20: Mild	3 (2.5)	2 (2.2)	2 (2.2)	3 (2.4)	2 (2.1)	2 (2.4)	3 (2.2)	1 (1.8)	2 (2.0)
20-29: Moderate	6 (5.8)	5 (5.1)	5 (5.1)	4 (4.9)	5 (4.2)	5 (4.9)	6 (5.1)	3 (4.1)	5 (4.8)
29-63: Severe	8 (8.7)	8 (8.7)	8 (7.7)	8 (7.7)	6 (6.7)	8 (7.7)	7 (8.7)	9 (7.1)	8 (8.2)
Age (years)									
20 and below	7 (5.1)	0 (4.5)	3 (4.5)	7 (4.5)	3 (3.9)	3 (4.5)	7 (4.7)	3 (3.8)	3 (4.4)
21-24	7 (8.7)	4 (7.7)	10 (7.7)	5 (7.7)	7 (6.7)	10 (7.7)	6 (8.0)	6 (6.5)	10 (7.5)
25-29	3 (2.9)	7 (2.6)	2 (2.6)	3 (2.4)	2 (2.1)	2 (2.4)	3 (2.9)	3 (2.4)	2 (2.7)
30 and above	0 (.4)	3 (.4)	0 (.3)	0 (.3)	1 (.3)	0 (.3)	0 (.4)	1 (.3)	0 (.3)
Language									
Mandarin	6 (5.1)	4 (4.5)	4 (4.5)	5 (4.2)	3 (3.6)	4 (4.2)	6 (4.7)	3 (3.8)	4 (4.4)
Cantonese	11 (11.9)	11 (10.5)	11 (10.5)	10 (10.8)	10 (9.4)	11 (10.8)	10 (11.3)	10 (9.2)	11 (10.6)

Note. Expected frequencies for each cell.

Instead, Monte Carlo simulations were used to assess random assignment adequacy across *Gender*, *Language*, *Age Range*, and *Depression Level*. Results (Table 5.19) indicated no significant associations for *Gender* by **Medium**, $p = 1.000$ (95% CI [1.000, 1.000]); *Language* by **Medium**, $p = .785$ (95% CI [.777, .793]); *Depression Level* by **Medium**, $p =$

1.000 (95% CI [1.000, 1.000]); and *Age Range* by **Medium**, $p = .722$ (95% CI [.713, .731]), suggesting successful random assignment.

Table 5. 19 Chi-Square Value on Age Range, Gender, Language, and Depression Level Across Medium in Study 1

	Value	*df*	Asymptotic Significance (2-sided)	Pearson Chi-square Assumption *	Monte Carlo Sig. (2-sided)		
					Significance	95% Confidence Interval	
						Lower Bound	Upper Bound
Gender × Medium	.197	2	.906	Violated (50.0%)	1.000	1.000	1.000
Language × Medium	.576	2	.750	Violated (50.0%)	.785	.777	.793
Depression Level × Medium	.438	4	.970	Violated (33.3%)	1.000	1.000	1.000
Age Range × Medium	4.325	6	.633	Violated (66.7%)	.722	.713	.731

Note. The assumption concerns the percentage of cells that have an expected count of less than five.

Subsequent Monte Carlo simulations corresponded to each timepoint with outliers excluded (*n1* = 47, *n2* = 43, *n3* = 44). Table 5.20 presents the results. Table 5.20 presents the results, revealing no significant correlations for any demographic variable and **Medium** at *T1*, *T2*, or *T3* (all $p > .05$). These results further support successful random assignment, with no significant demographic differences by **Medium** at any timepoint.

Table 5. 20 Chi-Square Value on Age Range, Gender, Language, and Depression Level Across Medium Without Outliers at Different Timepoints in Study 1

	Value	*df*	Asymptotic Significance (2-sided)	Pearson Chi-square Assumption *	Monte Carlo Sig. (2-sided)		
					Significance	95% Confidence Interval	
						Lower Bound	Upper Bound
				*n*1 = 47			
Gender × Medium	.186	2	.911	Violated (50.0%)	1.000	1.000	1.000
Language × Medium	.386	2	.824	Violated (33.3%)	.848	.841	.855
Depression Level × Medium	.231	4	.994	Violated (33.3%)	1.000	1.000	1.000
Age Range × Medium	4.704	6	.582	Violated (66.7%)	.652	.642	.661
				*n*2 = 43			
Gender × Medium	.439	2	.803	Violated (50.0%)	.907	.901	.913
Language × Medium	.382	2	.826	Violated (50.0%)	.917	.907	.918
Depression Level × Medium	.607	4	.962	Violated (66.7%)	.980	.977	.982
Age Range * Medium	6.256	6	.395	Violated (75.0%)	.401	.392	.411
				*n*3 = 44			
Gender × Medium	.111	2	.946	Violated (50.0%)	1.000	1.000	1.000
Language × Medium	.807	2	.668	Violated (50.0%)	.706	.697	.714
Depression Level × Medium	1.990	4	.738	Violated (55.6%)	.772	.764	.780
Age Range × Medium	5.864	6	.439	Violated (75.0%)	.453	.444	.463

Note. The assumption concerns the percentage of cells that have an expected count of less than five.

5.3.2 Descriptive Statistics of Experiential Variables Across Medium and Time

Table 5.21 presents descriptive statistics for eight experiential variables across **Medium** (*Audio*, *Video*, *Text*) at three different **Time** (*T1*, *T2*, *T3*), with and without outliers. Each cell shows the mean followed by the standard deviation in parentheses, and the number of observations (Num) for each medium at each time point is also reported at the bottom. Cells in bold indicate the highest value across mediums, an asterisk (*) indicates a higher value than the previous timepoint(s), and a double asterisk (**) indicates a higher value than the previous two timepoints.

With outliers, at *T1*, *Video* had the highest means for most variables, suggesting a better initial experience potentially due to movement and appearance cues. However, *Audio* had the highest *M_AffEase1* and *M_TrustP1*, indicating clear perception of interaction affordance and trust. At *T2*, Audio began to take precedence, with the highest mean for all variables except *M_ValEase2* where it tied with *Video*. This transition from initial visual preference to auditory preference may be due to ease of processing speech or reduced cognitive load. By *T3*, results were more mixed, with *Video* regaining higher evaluation on some variables (*M_ValEmo3*, *M_UseIntP3*, and *M_UseP_ci3*) but not others (*M_ValEase3*). Notably, *Audio* maintained the highest *M_AffEase* and **M_TrustP** throughout with outliers. *Text* started lower but showed an upward trend on *M_ValEmo*, *M_UseIntP*, *M_TrustP* and *M_UseP_ci*, indicating a potential adaptation effect over time.

Without outliers, trends shifted slightly, with *Video* most often having the highest values. *Audio* maintained the highest *M_AffEase* except at *T2* and highest *M_TrustP* at *T3*. *Text* showed similar trends, with slight value changes.

These are descriptive tendencies only without considering data structure, e.g., within-effect of repeated measurements and correlations. Also, the mean differences were subtle

except for *M_UseIntP1* and *M_TrustP1* across **Medium**; mean difference analyses are needed to determine significance.

Table 5. 21 Descriptive Statistics of Experiential Variables Across Medium and Time in Study 1

	T1			***T2***			***T3***		
Measures	Audio	Video	Text	Audio	Video	Text	Audio	Video	Text
				With outliers					
M_AffEase	**5.24 (.87)**	5.00 (.58)	4.89 (1.11)	**5.16 (.87)**	4.95 (.53)	4.95 (.94) *	**5.03 (.96)**	4.83 (.68)	4.73 (1.20)
M_RelEase	4.59 (1.05)	**4.63 (.64)**	4.58 (1.02)	**4.78 (.92)** *	4.64 (.69) *	4.56 (1.05)	**4.75 (1.12)**	4.72 (.54) **	4.43 (1.17)
M_ValEase	4.59 (.70)	**4.60 (.66)**	4.40 (1.22)	4.69 (.89) *	**4.69 (.53)** *	4.50 (1.24) *	**4.68 (.92)**	4.56 (.62)	4.46 (1.19)
M_ValEmo	4.19 (.90)	**4.36 (.56)**	4.00 (1.12)	**4.43 (.92)** *	4.36 (.83)	4.25 (1.08) *	4.34 (1.13)	**4.52 (.65)** *	4.44 (1.04) **
M_UseIntP	4.27 (.66)	**4.60 (.57)**	3.79 (.85)	**4.41 (.85)** *	4.33 (.67)	4.17 (1.28) *	4.20 (1.17)	**4.38 (.65)** *	4.29 (1.18) **
M_TrustP	**4.67 (1.03)**	4.31 (.73)	4.09 (.93)	**4.71 (.97)** *	4.44 (.83) *	4.34 (1.12) *	**4.56 (1.00)**	4.37 (.83)	4.37 (1.12) **
M_UseP_i	4.50 (1.02)	**4.75 (.48)**	4.62 (.94)	**4.76 (.94)** *	4.72 (.58)	4.47 (1.13) *	**4.62 (1.31)**	4.47 (.92)	4.47 (1.09)
M_UseP_ci	4.59 (1.15)	**4.66 (.94)**	4.41 (1.04)	**4.79 (1.00)** *	4.75 (.77) *	4.56 (1.29) *	4.67 (1.25)	**4.81 (.63)** **	4.63 (1.18) **
Num	17	16	16	17	16	16	17	16	16
				Without outliers					
M_AffEase	**5.23 (.87)**	4.98 (.60)	5.08 (.83)	**5.05 (.87)**	5.04 (.56) *	5.08 (.81)	**5.09 (.95)** *	4.92 (.59)	4.92 (.99)
M_RelEase	4.59 (1.05)	**4.62 (.66)**	4.75 (.77)	4.62 (.85) *	**4.87 (.39)**	4.73 (.83)	4.67 (1.10) **	**4.73 (.55)**	4.60 (1.00)
M_ValEase	4.59 (.70)	**4.64 (.66)**	4.56 (1.08)	4.58 (.88)	**4.79 (.50)** *	4.67 (1.08) *	4.67 (.95) *	**4.72 (.47)**	4.62 (1.03)
M_ValEmo	4.19 (.90)	**4.33 (.56)**	4.13 (1.02)	4.25 (.82) *	**4.48 (.66)** *	4.40 (.93) **	4.47 (1.02) **	**4.60 (.63)** **	4.60 (.84) **
M_UseIntP	4.27 (.66)	**4.58 (.58)**	3.87 (.82)	4.20 (.65)	**4.33 (.58)**	4.27 (1.26) *	4.40 (.85) *	**4.49 (.59)** *	4.42 (1.09) **
M_TrustP	**4.68 (1.03)**	4.33 (.75)	4.17 (.92)	4.53 (.90) *	**4.57 (.86)** *	4.47 (1.04) *	**4.69 (.87)** *	4.38 (.68)	4.53 (.95) **
M_UseP_i	4.50 (1.02)	**4.73 (.50)**	4.60 (.97)	4.63 (.92) *	**4.88 (.51)** *	4.63 (.95) *	**4.81 (1.06)** **	4.69 (.48)	4.63 (.90)
M_UseP_ci	4.59 (1.15)	**4.83 (.65)**	4.53 (.93)	4.67 (.99) *	**5.03 (.48)** *	4.73 (1.13) *	**4.97 (.96)** **	4.81 (.48)	4.80 (.98) **
Num	17	15	15	15	13	15	16	13	15

Note. Mean (SD) in each cell. Highest mean across Medium was in bold. An asterisk (*) indicates a higher value than the previous timepoint(s), and a double asterisk (**) indicates a higher value than the previous two timepoints.

5.3.3 Correlation Between Experiential Variables and Demographics

Before proceeding with mean difference analyses, it is critical to consider controlling for covariates like *Gender*, *Age Range*, *Language* and *Depression Level*. The general inclusion rule is relevance to the dependent variables (**DVs**). If a covariate is unrelated to any **DVs**, with no significant variation across groups, its inclusion would not change the results but unnecessarily complicate the model (Field, 2014). Per the demographic section, there

were no significant differences in demographics across **Medium**. Moreover, Table 5.22 showed no significant correlation between demographics and experiential variables, except for *Language*.

Given balanced *Language* distribution across **Medium** and no other significant covariate, the risk of a multicollinearity effect is mitigated (Tabachnick & Fidell, 2013). To streamline analysis and avoid unnecessary complexity, all demographic variables (*Gender*, *Age Range*, *Language*, *Depression Level*) were excluded as covariates in the subsequent mean difference analyses.

Table 5. 22 Correlation Between Demographic Variables and Experiential Variables in Study 1

	With outliers				Without outliers			
	Gender	*Language*	*AgeRange*	*DepLevel*	*Gender*	*Language*	*AgeRange*	*DepLevel*
M_AffEase1	.113 (.439)	**-.357 (.027)**	-.174 (.231)	-.020 (.890)	.176 (.237)	**-.367 (.011)**	-.078 (.601)	.046 (.760)
M_RelEase1	.222 (.125)	**-.348 (.014)**	-.148 (.312)	-.153 (.295)	.281 (.056)	**-.346 (.017)**	-.066 (.659)	-.116 (.438)
M_ValEase1	.178 (.222)	**-.357 (.012)**	-.121 (.409)	-.121 (.407)	.243 (.100)	**-.347 (.017)**	-.058 (.699)	-.065 (.663)
M_ValEmo1	.029 (.843)	**-.410 (.003)**	.000 (.997)	.217 (.133)	.054 (.716)	**-.418 (.003)**	.100 (.505)	-.201 (.175)
M_UseIntP1	.138 (.344)	-.270 (.061)	-.075 (.607)	-.010 (.946)	.160 (.284)	-.273 (.063)	.012 (.938)	.009 (.950)
M_TrustP1	.083 (.573)	-.265 (.066)	-.090 (.538)	-.077 (.601)	.108 (.471)	-.248 (.093)	-.060 (690)	-.044 (.768)
M_UseP1_i	.138 (.343)	**-.315 (.028)**	.131 (.371)	-.079 (.589)	.129 (.387)	**-.331 (.023)**	.134 (.369)	-.097 (.515)
M_UseP1_ci	.075 (.610)	**-.366 (.010)**	.085 (.561)	-.193 (.185)	.144 (.335)	**-.349 (.016)**	.097 (.518)	-.128 (.391)
M_AffEase2	.189 (.194)	**-.426 (.002)**	-.118 (.421)	-.017 (.905)	.244 (.114)	**-.460 (.002)**	-.064 (.681)	.101 (.521)
M_RelEase2	-.053 (.717)	**-.294 (.040)**	-.111 (.447)	-.164 (.261)	-.014 (.928)	**-.358 (.018)**	-.036 (.820)	-.021 (.892)
M_ValEase2	-.006 (.966)	**-.360 (.011)**	-.055 (.709)	-.131 (.368)	.027 (.866)	**-.385 (.011)**	-.011 (.943)	-.042 (.791)
M_ValEmo2	.017 (.909)	**-.412 (.003)**	-.114 (.435)	-.241 (.096)	.040 (.800)	**-.394 (.009)**	-.168 (.281)	-.170 (.275)
M_UseIntP2	.019 (.897)	**-.306 (.033)**	-.109 (.457)	-.024 (.868)	-.006 (.969)	-.269 (.081)	-.131 (.404)	.050 (.751)
M_TrustP2	-.152 (.296)	-.234 (.105)	-.085 (.564)	-.132 (.367)	-.161 (.302)	-.241 (.120)	-.031 (.842)	-.029 (.852)
M_UseP2_i	-.182 (.212)	**-.296 (.039)**	-.057 (.695)	-.076 (.603)	-.178 (.254)	**-.320 (.037)**	.013 (.934)	.053 (.737)
M_UseP2_ci	-.157 (.282)	**-.362 (.010)**	-.072 (.624)	-.247 (.087)	-.127 (.415)	**-.392 (.009)**	-.062 (.694)	-.139 (.375)
M_AffEase3	.261 (.071)	**-.435 (.002)**	-.080 (.586)	.001 (.997)	.309 (.041)	**-.483 (<.001)**	-.011 (.945)	.011 (.941)
M_RelEase3	254 (.078)	**-.296 (.039)**	-.076 (.604)	-.062 (.673)	.278 (.068)	**-.308 (.042)**	-.009 (.956)	-.069 (.658)
M_ValEase3	.140 (.338)	**-.296 (.039)**	-.130 (.374)	-.098 (.502)	.193 (.208)	**-.375 (.012)**	-.072 (.642)	-.126 (.417)
M_ValEmo3	.089 (.543)	**-.384 (.006)**	-.145 (.319)	-.094 (.519)	.196 (.202)	**-.453 (.002)**	-.067 (.665)	-.025 (.870)
M_UseIntP3	.039 (.788)	-.150 (.303)	-.122 (.402)	-.012 (.934)	.126 (.414)	-.159 (.302)	-.072 (.640)	.070 (.649)
M_TrustP3	.144 (.324)	-.228 (.115)	-.219 (.131)	-.098 (.503)	.225 (.141)	-.194 (.053)	-.141 (.362)	-.050 (.746)
M_UseP3_i	.167 (.250)	**-.362 (.011)**	-.010 (.943)	-.036 (.806)	.161 (.266)	**-.396 (.008)**	.005 (.973)	-.090 (.562)
M_UseP3_ci	.099 (.499)	-.272 (.058)	-.199 (.170)	-.163 (.263)	.238 (.120)	**-.347 (.021)**	-.106 (.492)	-.078 (.614)

5.3.4 Generalized Estimating Equations on Experiential Variables

5.3.4.1 Effects of Medium on Experiential Variables across Time

Generalized Estimating Equations (GEE) were employed to assess the impact of **Time** (*T1* versus *T2* versus *T3*) and **Medium** (*Audio* versus *Video* versus *Text*) on participant experiences: *M_AffEase*, *M_RelEase*, *M_ValEase*, *M_ValEmo*, *M_UseIntP*, *M_TrustP*, *M_UseP_i*, *M_UseP_ci*. Pairwise comparisons identified changes between specific timepoints and **Medium**. GEE was chosen due to non-normal distribution of the experiential variables, confirmed by the Shapiro-Wilk tests (Table 5.23), and the need to account for within-subject correlation.

Table 5. 23 Test of Normality on Experiential Variables for T1, T2, and T3 Using Shapiro-Wilk Tests (N = 49) in Study 1

	Medium	***T1***			***T2***			***T3***		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.	Statistic	*df*	Sig.
M_AffEase	Audio	.825	17	.005	.854	17	.012	.799	17	.002
	Video	.960	16	.670*	.939	16	.341*	.954	16	.548*
	Text	.880	16	.039	.917	16	.149*	.885	16	.047
M_RelEase	Audio	.904	17	.079*	.931	17	.228*	.900	17	.069*
	Video	.851	16	.014	.809	16	.004	.836	16	.008
	Text	.921	16	.177*	.935	16	.287*	.919	16	.161*
M_ValEase	Audio	.866	17	.019	.935	17	.268*	.923	17	.164*
	Video	.959	16	.639*	.854	16	.016	.756	16	<.001
	Text	.946	16	.429*	.909	16	.111*	.934	16	.283*
M_ValEmo	Audio	.910	17	.100*	.949	17	.438*	.920	17	.150*
	Video	.924	16	.192*	.924	16	.194*	.980	16	.962*
	Text	.965	16	.753*	.952	16	.515*	.958	16	.621*
M_UseIntP	Audio	.946	17	.401*	.917	17	.134*	.928	17	.202*
	Video	.888	16	.052	.855	16	.016	.827	16	.006
	Text	.937	16	.309*	.901	16	.085*	.926	16	.209*
M_TrustP	Audio	.912	17	.109*	.906	17	.087*	.906	17	.084*
	Video	.926	16	.211*	.898	16	.075*	.914	16	.137*
	Text	.922	16	.181*	.929	16	.232*	.948	16	.453*
M_UseP_i	Audio	.947	17	.412*	.926	17	.185*	.869	17	.021
	Video	.737	16	<.001	.787	16	.002	.611	16	<.001
	Text	.918	16	.155*	.937	16	.312*	.925	16	.204*
M_UseP_ci	Audio	.898	17	.063*	.889	17	.045	.849	17	.010
	Video	.870	16	.027	.883	16	.043	.818	16	.005
	Text	.905	16	.097*	.892	16	.060*	.906	16	.102*

Note. This is a lower bound of true significance.

The analytic approach began with GEEs applied to the complete dataset with outliers ($N = 49$), no missing data. Two models were devised: **Model 1** analyzed effects of **Time**, **Medium**, and their interaction (**Time** × **Medium**) on the dependent variables; if non-significant, **Model 2** reassessed just **Time** and **Medium** main effects.

SPSS 28 GEE models for each dependent variable included **Medium** as fixed factor and **Time** as within-factors. Gamma distribution with an identity link function and an unstructured working correlation matrix were used to account for skewness and different rates of change between timepoints, respectively.

Model 1 results (Table 5.24, left side) showed a significant interaction only for *M_UseIntP* ($\chi 2 = 19.98$, $df = 4$, $p < .001$). Thus, **Model 2** was run for the other seven variables. **Model** results (Table 5.24, right side) revealed a significant main **Time** effect on *M_ValEmo* in **Model 1** ($\chi 2 = 6.17$, $df = 2$, $p = .046$**)** and near significance in **Model 2** ($\chi 2 = 6.17$, $df = 2$, $p = .046$**)**. No other significant effects emerged.

Table 5. 24 Tests of Generalized Estimating Equation Model Effects on Experiential Variables (Model 1: With Interaction Effect and Model 2: With Interaction Effect) with N = 49 in Study 1

	Model 1: Type III			**Model 2: Type III**		
Experiential variables	Wald Chi-Square	*df*	*Sig.*	Wald Chi-Square	*df*	*Sig.*
M_AffEase						
Time	5.013	2	.082	5.103	2	.078
Medium	1.138	2	.566	1.198	2	.549
Medium × Time	.615	4	.961	/	/	/
M_RelEase						
Time	.404	2	.817	.433	2	.805
Medium	.329	2	.848	.319	2	.852
Medium × Time	1.650	4	.800	/	/	/
M_ValEase						
Time	1.309	2	.520	1.321	2	.517
Medium	.371	2	.831	.386	2	.825
Medium × Time	.431	4	.980	/	/	/
M_ValEmo						
Time	6.170	2	.046	5.918	2	.052
Medium	.500	2	.779	.614	2	.736
Medium × Time	3.665	4	.453	/	/	/

M_UseIntP						
Time	1.075	2	.584	/	/	/
Medium	1.766	2	.414	/	/	/
Medium × Time	19.980	4	<.001	/	/	/
M_TrustP						
Time	2.249	2	.325	2.129	2	.345
Medium	1.764	2	.414	1.879	2	.391
Medium × Time	1.526	4	.822	/	/	/
M_UseP_i						
Time	1.537	2	.464	1.478	2	.477
Medium	.237	2	.888	1.142	2	.565
Medium × Time	3.325	4	.505	/	/	/
M_UseP_ci						
Time	1.985	2	.371	1.957	2	3.76
Medium	.482	2	.786	.521	2	.771
Medium × Time	.298	4	.990	/	/	/

Pairwise comparisons (Table 5.25) showed higher *M_ValEmo* scores in *T3* versus *T1* within **Model 1** ($MD_{(T3-T1)} = .249$, $p = .042$) and near significance in **Model 2** ($MD_{(T3-T1)} = .246$, $p = .048$). The *M_UseIntP* interaction effects arose from **Medium** (*Video* versus *Text*) in *T1* ($MD_{(video - text)} = .812$, $p = .039$) and **Time** (*T3* versus *T1*) with all participants ($MD_{(T3 - T1)} = .249$, $p = .042$), indicating lower initial use intention (*M_UseIntP*) for *Text* initially but gradual willingness over time. *Video* also had higher *M_UseIntP* than *Text* in *T1*.

Table 5. 25 Pairwise Comparisons on M_ValEmo and M_UseIntP with Model 1 and Model 2 in Study 1

Condition	Means (SE)*	Mean Difference	Std. Error	95% Wald Confidence Interval [Lower, Upper]	Sig
		M_ValEmo (within Model 1)			
T3	4.428 (.95)	.2489	.101118	.0067, .4911	.042
T1	4.184 (.88)				
		M_ValEmo (within Model 2)			
T3	4.428 (.95)	.2462	.10206	.0018, 4905	.048
T1	4.184 (.88)				
		M_UseIntP (within Model 1)			
Video & T1	4.60 (.57)	.8125	.24842	.0183, 1.6067	.039
Text & T1	3.79 (.85)				
Text & T3	4.20 (1.18)	.5000	.14130	.0483, .9517	.014
Text & T1	3.79 (.85)				

Note. The means reported for the GEE models represent model-estimated marginal means that account for correlations and other model parameters like outlier removal. In contrast, the ANOVA provides unweighted raw means. The GEE estimated means can more accurately reflect the true population average response.

GEEs were not run on no-outlier datasets due to insufficient samples (n = 39) and altered distribution from outlier removal, generating new outliers. GEE inherently handles outliers methodologically.

In summary, no significant **Medium** or **Time** effects occurred, except for **Medium** on *M_UseIntP1* and **Time** on *M_ValEmo* from the *Text* group. To further examine whether participants experienced **eiIBM_RobotV1** similarly across **Medium**, data were analyzed in JASP using repeated Bayesian ANOVA (van den Bergh et al., 2022)

5.3.5 Bayesian Analyses on Experiential Variables

In exploring human-robot interaction nuances, Hoorn and Winter (2017) proposed two contrasting hypothesis tests: the similarity hypothesis assumes people respond to robots akin to human counterparts, adhering to social norms; the dissimilarity hypothesis suggests subtle deviations due to robots' non-human nature. Conventional (frequentist) techniques like GEE test against a null hypothesis (H0), indicating dissimilarity when significant differences emerge. However, non-significant differences do not confirm similarity, only a lack of observed difference.

The prior GEEs did not find significant experiential differences between-subject **Medium** or with-subject *Time*, except in specific cases. Thus, identical experiences and perceptions across **Medium** could not be claimed, highlighting conventional statistics' limitation in confirming similarity. Nor could significant findings be confidently generalized, as these methods do not incorporate priori knowledge.

Inspired by Hoorn and Winter (2017), Bayesian techniques were applied to analyze experiential data. Bayesian analysis integrated prior knowledge and observed data into a posterior distribution reflecting updated understanding of variables of interest.

5.3.5.1 Bayesian Repeated Measures ANOVAs

Using JASP, Bayes Factors (BF) evaluated hypothesis evidence. Eight repeated measures ANOVAs were conducted per variable across 3 timepoints ($N = 49$). JASP calculated each model's BF relative to a null model; BF10 indicates evidence factoring the alternative hypothesis. Given the novel I-PEFiC application in robot-delivered intervention, it lacked prior knowledge of people's experiences. Thus, a uniform .20 prior probability was set for each model and BF10 of 1.000 as the null baseline likelihood.

BF interpretations followed Hoorn and Winter (2017) based on Jeffreys (1961): BF10 3-30 suggests moderate to strong H1 evidence; BF10 1/30 – 1/3 indicates moderate to strong H0 evidence. Bayesian repeated measures ANOVAs revealed cross-**Medium** similarities at three timepoints (Table 5.26).

Results consistently favored the null across measures (BF10 range: 0.002-0.734), strengthening in more complex models. Posterior probabilities indicated negligible contributions of **Time**, **Medium**, and their interaction (P (incl | data) ≤ 0.301). Thus, experiential scores appear unexplained by **Time** or **Medium**, suggesting experience similarity.

Table 5. 26 Bayesian Repeated Measures ANOVA Results for Experiential Variables with N = 49 in Study 1

Measure		Models				
		Null	**Medium**	**Time**	**Time + Medium**	**Time + Medium + Time × Medium**
M_AffEase	BF_{10}	1.000	.438	.391	.179	.010
	Evidence strength to H0	/	Slight	Slight	Moderate	Very strong
M_RelEase	BF_{10}	1.000	.287	.078	.023	.002
	Evidence strength to H0	/	Moderate	Moderate	Strong	Very strong
M_ValEase	BF_{10}	1.000	.350	.104	.038	.002
	Evidence strength to H0	/	Slight	Moderate	Moderate	Very strong
M_ValEmo	BF_{10}	1.000	.734	.295	.216	.026
	Evidence strength to H0	/	Slight	Moderate	Moderate	Strong
M_UseIntP	BF_{10}	1.000	.393	.090	.053	.033
	Evidence strength to H0	/	Slight	Moderate	Moderate	Strong
M_TrustP	BF_{10}	1.000	.443	.125	.056	.006
	Evidence strength to H0	/	Slight	Moderate	Moderate	Very strong
M_UseP_i	BF_{10}	1.000	.230	.112	.027	.003

	Evidence strength to H0	/	Moderate	Moderate	Strong	Very strong
M_UseP_ci	BF_{10}	1.000	.271	.189	.055	.002
	Evidence strength to H0	/	Moderate	Moderate	Moderate	Very strong

Note. "Slight" for BF10 values that do not provide moderate evidence for H0 (greater than 0.334 but less than 1); "Moderate" for BF10 values between 0.334 and 0.033; "Strong" for BF10 values between 0.033 and 0.010; "Very Strong" for BF10 values less than 0.010.

5.3.5.2 Independent and paired comparison

Despite overall null findings, further analysis investigated specific GEE-identified effects: **Time** on valence of emotional distress relief (*M_ValEmo*) between *T1* and *T3*; **Medium** on use intention (*M_UseIntP1*) between *Video* and *Text* at *T1*. For consistency and to account for non-normality, Bayesian Wilcoxon Signed-Rank and Mann-Whitney U Tests were used.

The Bayesian Wilcoxon Signed-Rank Test provided Bayes Factors (BF_{10}) indicating the strength of evidence for the alternative hypothesis over the null. Table 5.27 summarized the results. For *M_ValEmo*, *T1* versus *T3* yielded BF10 of 2.115, indicating slight evidence for a small, significant difference, aligning with GEE results (p = 0.48 in **Model 1** and p = 0.52 in **Model 2**). Additionally, *M_AffEase2* versus *M_AffEase3* produced BF10 of 1.255, exceeding 1 and marginally indicating a difference in affordance of ease-of-use interaction (p = 0.88 in **Model 1** and p = 0.85 in **Model 2**). This modest evidence may be insufficient for definitive conclusions, consistent with the non-significant GEE findings. Other **Time** comparisons favored the null, reflecting experience similarity.

Table 5. 27 Bayesian Wilcoxon Signed-Rank Test (BF^{10} Value in Each Cell) with N = 49 in Study 1

Measure 1		Measure 2	BF_{10}	W	Rhat
M_AffEase1	-	M_AffEase2	0.154	309.000	1.000
M_AffEase1	-	M_AffEase3	0.362	320.000	1.001
M_AffEase2	-	M_AffEase3	1.255	354.500	1.001
M_RelEase1	-	M_RelEase2	0.189	320.500	1.000
M_RelEase1	-	M_RelEase3	0.162	262.000	1.000

M_RelEase2	-	M_RelEase3	0.163	204.000	1.000
M_ValEase1	-	M_ValEase2	0.266	280.500	1.001
M_ValEase1	-	M_ValEase3	0.159	262.000	1.000
M_ValEase2	-	M_ValEase3	0.213	249.500	1.001
M_ValEmo1	-	M_ValEmo2	0.526	288.000	1.000
M_ValEmo1	-	M_ValEmo3	2.115	248.000	1.000
M_ValEmo2	-	M_ValEmo3	0.215	290.500	1.000
M_UseIntP1	-	M_UseIntP2	0.268	250.000	1.001
M_UseIntP1	-	M_UseIntP3	0.221	355.000	1.000
M_UseIntP2	-	M_UseIntP3	0.160	376.000	1.000
M_TrustP1	-	M_TrustP2	0.305	109.000	1.000
M_TrustP1	-	M_TrustP3	0.229	193.500	1.000
M_TrustP2	-	M_TrustP3	0.217	154.000	1.000
M_UseP1_i	-	M_UseP2_i	0.176	195.500	1.000
M_UseP1_i	-	M_UseP3_i	0.167	187.500	1.001
M_UseP1_ci	-	M_UseP2_ci	0.313	227.000	1.000
M_UseP1_ci	-	M_UseP3_ci	0.468	208.500	1.002
M_UseP2_ci	-	M_UseP3_ci	0.186	142.500	1.001

Similarly, the Bayesian Mann-Whitney U Text compared the effects of **Medium** on the experiential measures across different timepoints (Table 5.28). Comparing *M_UseIntP* between *Video* and *Text* at *T1* yielded BF10 of 4.429 with outliers, indicating moderate **Medium** effect evidence aligning with GEE. However, excluding one *Video* and *Text* outlier reduced BF10 to 2.306, implying the effect depends more on the individual experience than **Medium**.

Table 5. 28 Bayesian Mann-Whitney U Test (BF^{10} Value in Each Cell) With and Without Outliers in Study 1

Time	Comparisons	Measure							
		M_AffEase	*M_RelEase*	*M_ValEase*	*M_ValEmo*	*M_UseIntP*	*M_TrustP*	*M_UseP1_i*	*M_UseP1_ci*
					Audio versus *Video*				
T1	*Audio (17) versus Video (16)*	.549	.351	.335	.357	.689	.526	.366	.346
	Audio (17) versus Video (15)	.610	.334	.361	.355	.592	.501	.375	.384
T2	*Audio (17) versus Video (16)*	.517	.374	.355	.334	.346	.454	.57	.362
	Audio (15) versus Video (13)	.384	.449	.383	.456	.416	.369	.477	.497
T3	*Audio (17) versus Video (16)*	.445	.379	.385	.380	.348	.371	.408	.362
	Audio (16) versus Video (13)	.436	.382	.377	.399	.380	.502	.412	.400
					Audio versus *Text*				
T1	*Audio (17) versus Text (16)*	.447	.340	.392	.361	1.000	.828	.368	.395
	Audio (17) versus Text (15)	.372	.394	.368	.344	.666	.637	.348	.365
T2	*Audio (17) versus Text (16)*	.434	.389	.358	.383	.449	.442	.349	.378
	Audio (15) versus Text (15)	.359	.379	.385	.391	.363	.363	.365	.363
T3	*Audio (17) versus Text (16)*	.435	.423	.375	.332	.350	.364	.382	.373
	Audio (16) versus Text (15)	.410	.365	.360	.381	.344	.372	.399	.407
					Video versus *Text*				

T1	*Video (16) versus Text (16)*	.347	.344	.380	.612	4.429	.430	.380	.399
	Video (15) versus Text (15)	.368	.382	.369	.464	2.306	.401	.371	.478
T2	*Video (16) versus Text (16)*	.368	.333	.351	.339	.372	.353	.399	.356
	Video (13) versus Text (15)	.374	.409	.376	.346	.383	.384	.436	.436
T3	*Video (16) versus Text (16)*	.333	.376	.353	.352	.356	.353	.336	.347
	Video (13) versus Text (15)	.384	.378	.378	.363	.366	.399	.381	.370

Note. Each comparison was performed with and without outliers in each dataset.

In summary, evidence from GEEs and Bayesian analysis generally favors perception similarity across **Medium** and **Time**, with indications of potential small differences in specific variables.

5.3.6 Path Analyses on Experiential Variables

5.3.6.1 Multi-Group Path Analyses on Experiential Variables

I utilized partial least squares path modeling (PLSPM) in SmartPLS 4 to investigate sensory information processing within the I-PEFiC framework for **eiIBM_RobotV1**. PLS-SEM was chosen due to its robustness with small sample sizes and non-normally distributed data. Increasing the sample size was not feasible within the thesis timeline.

To validate the theoretical framework and assess path consistency over time, I planned separate PLS-SEM models for each timepoint (*T1*, *T2*, *T3*) using multi-group analysis (MGA). Separate models by **Medium** ($n < 19$) were not feasible due to the unreliability of covariance-based SEM with small samples. Given the absence of significant **Medium** effects in prior GEEs and Bayesian analysis results, a holistic data analysis approach was reasonable.

The model in Figure 5.8, comprising 6 I-PEFiC constructs, was analyzed following the MICOM procedure (Henseler et al., 2016) to establish measurement invariance for valid MGA interpretation. Despite the challenges posed by small samples in PLS-SEM, I justified the adequacy of the model's low evaluated complexity. By including only key constructs and

relationships based on theory, model complexity was reduced, enhancing stability with the given sample size (Hair et al., 2018).

MGA assumes heterogeneity across groups, making it useful for assessing differences over **Time**. Establishing measurement invariance is crucial before MGA, as failure to do so can lead to poor statistical power and misleading results (Hult et al., 2008). Measurement invariance ensures consistent attribute measurement across conditions. The MICOM procedure systematically assesses measurement invariance through three key steps:1) configural invariance assessment, automatically confirmed in SmartPLS 4; 2) compositional invariance assessment by comparing correlations to 5% permutation quantiles, which were exceeded to establish compositional invariance; and 3) composite equality assessment by comparing original mean differences to 2.5% and 97.5% permutation boundaries. Results within boundaries and $p > .05$ indicate composite equality.

Achieving Step 1, Step 2, and either requirement of Step 3 (equality of composite variance or equality of composite mean) allows for partial measurement invariance and proceeding with MGA. If all three Steps are achieved, full measurement invariance can be claimed, rendering MGA unnecessary (Henseler et al., 2016). In such cases, pooling the data is a feasible option to increase statistical power (Cheah et al., 2020).

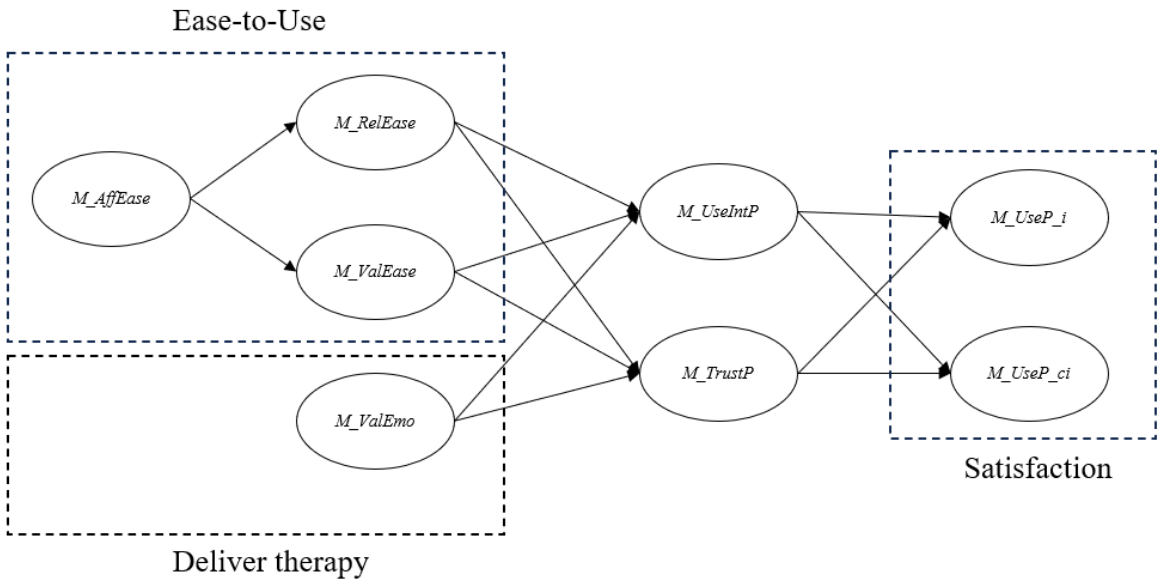

Figure 5.8 Partial I-PEFiC model for Study 1

I performed permutation multigroup analysis between *T1 - T2*, *T1 - T3*, and *T2* - T3 (Table 5.29) with 1,000 permutations and two-tailed .05 significance to assess compositional invariance and composite equality.

The *T1- T2* comparison showed a significant permutation *p*-value, indicating differing composite forms between timepoints and suggesting inconsistent constructs across *T1- T2*. Consequently, conducting multigroup analysis between these time periods would not be methodologically meaningful due to the lack of measurement invariance. Importantly, this finding aligns with and validates the PCA results, which also revealed *T1* data inconsistencies with the *T2* data.

In contrast, the *T1- T3* comparison met compositional invariance for most constructs, with the sole exception of *M_ValEmo*. However, the presence of all constructs within the 95% confidence interval of the permutation test boundaries indicates no significant differences in the mean values and variances of the latent variables between *T1* and *T3*. Therefore, the *T1* and *T3* data demonstrated full measurement invariance, rejecting the necessity of MGA between these two timepoints.

Similarly, the *T2- T3* results showed compositional non-invariance for three constructs: *M_AffEase*, *M_ValEase*, and *M_UseP1_i.* However, equal variance and means were confirmed, rejecting the need for MGA.

Table 5. 29 Measurement Invariance Assessment on Experiential Variables According to MICOM in Study 1

Measure	Compositional Invariance Assessment		Composite Equality Assessment					
			Mean Invariance			Variance Invariance		
	Original /5% Correlation	Permutation p-Value	Original mean difference	2.5%, 97.5%	Permutation p-Value	Original invariance difference	2.5%, 97.5%	Permutation p-Value
				T1 versus *T2*				
M_AffEase	1.00 / 1.00	.015	.025	-.383, .420	.919	.200	-.646, .623	.638
M_RelEase	1.00 / 1.00	.000	-.075	-.397, .385	.738	.037	-.609, .549	.908
M_ValEase	1.00 / 1.00	.000	-.107	-.396, .412	.611	.078	-.602, .588	.786
M_ValEmo	1.00 / 1.00	.105	-.181	-.396, .396	.381	-.102	-.552, .523	.689
M_TrustP	1.00 / 1.00	.011	-.141	-.379, .401	.502	-.107	-.443, .427	.606
M_UseIntP	1.00 / 1.00	.000	-.095	-.398, .413	.679	-.440	-.493, .528	.096
M_UseP1_i	1.00 / 1.00	.000	-.150	-.410, 410	.501	.013	-. 549, .546	.955
M_UseP1_ci	1.00 / 1.00	.000	-.035	-.390, 390	.896	-.147	-.582, .589	.631
				T1 versus *T3*				
M_AffEase	1.00 / 1.00	.260	.196	-.397, .408	.324	-.175	-.714, .628	.611
M_RelEase	1.00 / 1.00	.581	-.044	-.416, .383	.829	-.154	-.580, .566	.617
M_ValEase	1.00 / 1.00	.166	-.046	-.411, .410	.815	-.097	-.624, .591	.756
M_ValEmo	1.00 / 1.00	.052	-.267	-.379, .390	.193	-.146	-.528, .515	.561
M_TrustP	1.00 / 1.00	.597	-.076	-.402, .381	.697	-.103	-.480, .471	.674
M_UseIntP	1.00 / 1.00	.000	-.068	-.406, .406	.762	-.565	-.675, .721	.103
M_UseP1_i	1.00 / 1.00	.547	-.178	-.416, .416	.372	-.018	-.612, .627	.951
M_UseP1_ci	1.00 / 1.00	.278	.105	-.400, .400	.633	-.541	-.684, .701	.150
				T2 versus *T3*				
M_AffEase	1.00 / 1.00	.000	.181	-.392, .403	.380	-.375	-.618, .576	.245
M_RelEase	1.00 / 1.00	.334	.028	-.381, .392	.919	-.191	-.624, .604	.548
M_ValEase	1.00 / 1.00	.008	.060	-.403, .418	.763	-.019	-.642, .564	.953
M_ValEmo	1.00 / 1.00	.120	.088	-.405, .372	.656	-.045	-.555, .489	.876
M_TrustP	1.00 / 1.00	.011	.064	-.403, .381	.797	.004	-.532, .472	.989
M_UseIntP	1.00 / 1.00	.468	.022	-.426, .399	.905	-.125	-.568, 585	.701
M_UseP1_i	1.00 / 1.00	.326	-.030	-.409, .389	.909	-.031	-.621, .565	.923
M_UseP1_ci	1.00 / 1.00	.117	.133	-.399, .399	.531	-.394	-.713, .674	.269

5.3.6.2 Separate Path Analyses at T1, T2 and T3

Given the lack of partial measurement invariance between all timepoints, multigroup analysis was unnecessary. Instead, separate path analyses were conducted at each timepoints (*T1*, *T2*, *T3*) using SmartPLS for PLS-SEM to assess the impacts of human-robot interaction

over time. Bootstrap methods were used to estimate standardized path coefficients (β) and *p*-values. The results are depicted in Table 5.30.

Results showed a consistently significant positive path from *M_AffEase* to *M_RelEase* and *M_ValEase* across timepoints, with intensifying coefficients and *p*-values over time. Specifically, the path from *M_AffEase* to *M_RelEase* showed an increasing β coefficients of 0.58 at *T1* and 0.742 at *T3*, with *p*-values remaining at .000; the path from *M_AffEase* to *M_ValEase* demonstrated the β coefficients increased from 0.554 at *T1* to 0.731 at *T3*, also with *p*-values remaining at .000. However, the path from *M_RelEase* to *M_TrustP* was consistently non-significant (*p*s > .05), suggesting *M_RelEase* does not impact *M_TrustP* for **eiIBM_RobotV1**. *M_RelEase* also did not significantly affect *M_UseIntP* at any timepoint.

M_TrustP significantly predicted both *M_UseP_i* (β = 0.455, *p* = .000 at *T1*; β = 0.436, *p* = .000 at *T2*) and *M_UseP_ci* (β = 0.546 *p* = .000 at *T1*; β = 0.461, *p* = .000 at *T2*) at *T1* and *T2*, but only *M_UseP_i* at *T3* (β = 0.274, *p* = .024). *M_UseIntP* influence on *M_UseP_ci* and *M_UseP_i* grew increasingly significant over time (*M_UseP_ci*: β = 0.208, *p* = .160 at *T1* to β = 0.640, *p* = .000 at *T3; M_UseP_i*: β = 0.298, *p* = .021 at *T1* to β = 0.654, *p* = .000 at *T3*).

M_ValEase significantly impacted *M_TrustP* at *T1* (β = 0.349, *p* = .015) and *T2* (β = 0.499, *p* = .005) but became non-significant by *T3* (β = 0.234, *p* = .186). It did not significantly affect *M_UseIntP* over time. In contrast, *M_ValEmo* consistently and strongly predicted *M_UseIntP* (β = 0.485, *p* = .000 at *T1*; β = 0.566, *p* = .000 at *T1*; β = 0.766, *p* = .001 at *T3*), underscoring its emotional effect on use intention. *M_ValEmo* only significantly influenced *M_TrustP* at *T1*(β = 0.41, *p* = .002) and *T3* (β = 0.547, *p* = .006).

Table 5. 30 Path Analyses Results on Experiential Variables for T1, T2, and T3 in Study 1

Path	*T1*			*T2*			*T3*		
	β	*t*	*p*	β	*t*	*p*	β	*t*	*p*
M_AffEase -> M_RelEase	0.58	4.33	0	0.592	5.764	0	0.742	8.379	0

M_AffEase -> M_ValEase	0.554	4.434	0	0.634	6.919	0	0.731	8.79	0
M_RelEase -> M_TrustP	0.078	0.494	.621	0.103	0.517	.605	0.116	0.705	.481
M_RelEase -> M_UseIntP	0.129	1.095	.274	0.139	1.077	.281	0.01	0.051	.959
M_TrustP -> M_UseP_ci	0.546	5.466	0	0.461	3.011	.003	0.274	2.252	.024
M_TrustP -> M_UseP_i	0.455	4.78	0	0.436	3.145	.002	0.198	1.654	.098
M_UseIntP -> M_UseP_ci	0.208	1.404	.160	0.414	2.741	.006	0.64	5.048	0
M_UseIntP -> M_UseP_i	0.298	2.305	.021	0.379	2.845	.004	0.654	4.779	0
M_ValEase -> M_TrustP	0.349	2.438	.015	0.499	2.786	.005	0.234	1.322	.186
M_ValEase -> M_UseIntP	0.127	0.951	.341	0.221	1.416	.157	0.101	0.524	.600
M_ValEmo -> M_TrustP	0.41	3.157	.002	0.265	1.708	.088	0.547	2.743	.006
M_ValEmo -> M_UseIntP	0.485	3.611	0	0.566	4.461	0	0.766	3.223	.001

Although some construct inconsistencies emerged between *T1* and *T2*, overall path relationships remained relatively stable, suggesting that while individual constructs changed, the overarching morel was consistent.

Figure 5.9 visually compares the path analyses results at different timepoints. In summary, while *M_RelEase* and *M_ValEase* did not always directly impact *M_TrustP* and *M_UseIntP*, the persistent *M_AffEase* influence on *M_RelEase* and *M_ValEase* confirms I-PEFiC's significance over time. Critically, *M_ValEmo* strongly predicted *M_UseIntP*, highlighting the importance of emotional relief affordance to the depressed participants.

The results emphasize that beyond ease-of-use interaction, the robot's ability to alleviate emotional distress strongly motivates engagement with **eiIBM** robots therapeutically. According to the coefficients of *ValEase* and *ValEmo* to *UseIntP* and *TrustP*, it indicated that the valence of emotional relief contributed more to the use intention of the program. Emotional resonance emerges as a pivotal factor affecting acceptance and perceived effectiveness of **eiIBM_RobotV1**, potentially impacting overall treatment outcomes. The I-PEFiC framework helps explain this finding by filtering and revealing how valence and relevance contexts contribute to use intention and engagement with robotic agents.

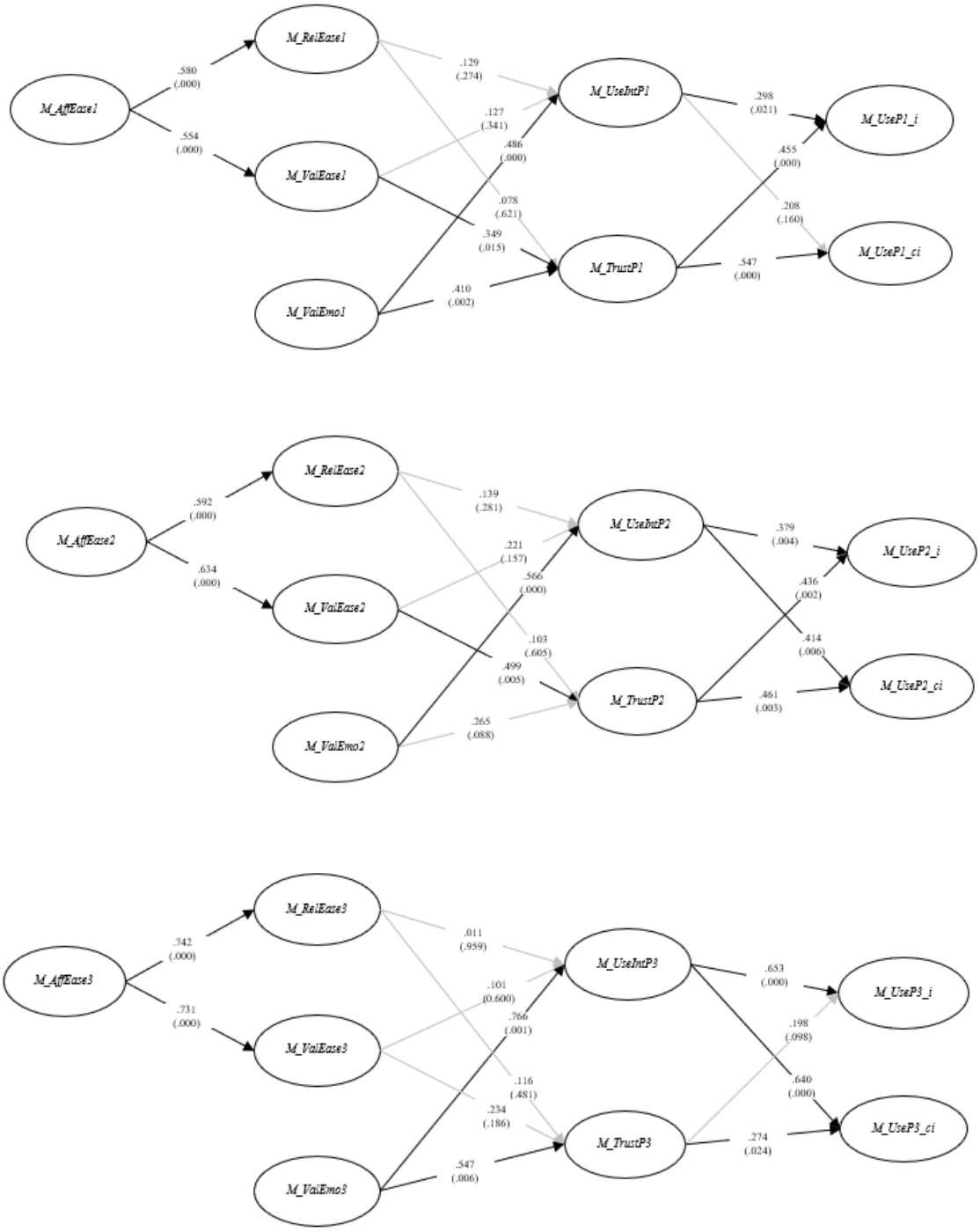

Figure 5.9 Upper: Path Analyses Results at T1; Middle: Path Analyses Results at T2; Bottom: Path Analyses Results at T3.

5.3.7 Effects of Medium on Intervention Outcome Over Time

Given the established validity and reliability of the assessments used (*Chapter 4*), re-evaluation of psychometric properties was unnecessary. Mean indicator scores for each assessment were calculated across groups and analyzed using GEE to examine changes over time.

For intervention outcome data, the focus was on interaction effects between **Medium (***Audio*, *Video*, *Text, Control***)** and **TestTime (***pretest*, *posttest***)** in GEE models. Pre-test scores represented baseline data, so pre-existing differences between **Medium** groups at baseline were not theoretically desired and post-existing differences were less interpretable without considering the baseline. Instead, analyses inspected whether the trajectory of change on intervention outcomes from pre- to post-intervention differed depending on the **Medium**. GEEs analyzed effects of **TestTime**, **Medium**, and their interaction (**TestTime** × **Medium**) on six assessment measures: *DS_MS*, *SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*. Pairwise comparisons identified changes among **Medium** groups.

GEE results (Table 5.31 and Table 5.32) showed significant main effects of **TestTime** for all outcome measures, indicating symptom improvements over the course of the **eiIBM_RobotV1** regardless of **Medium**. Significant **Medium** x **TestTime** interactions were also found, suggesting differential degrees of change depending on robot modality.

For *DS_MS*, the intervention groups showed pre-post decreases with medium-large effect sizes (ranging from 14.72 to 15.50, Hedge's g = 1.329 to 1.114), while *Control* was stable (3.72 change, *SE* = 0.15, 95% CI [-0.10, 0.84], *p* = .364), all *ps* =.000. Between-group comparisons revealed no significant differences at *pretest* or *posttest* after Bonferroni correction (*p*s > 0).

On *SST_TNR*, the intervention groups showed pre-post reductions of .208 to .264 with medium effect sizes (Hedge's g = 0.414 to 0.476), while *Control* was stable (6.3 change), all *ps* < 0.004. Groups did not differ significantly at *pretest* or *posttest* (*p*s = 1.000), according to pairwise comparison.

For *SRT_PT*, the intervention groups evidenced pre-post decreases of .637 to .924 with medium-large within-group effect sizes (Hedge's g = 0.549 to 0.713). *Control* remained stable (0.61 change). At *posttest*, *Audio* and *Text* differed significantly from *Control* with medium-large between-group effect sizes (Hedge's g = 0.710 and 0.604, respectively).

On *SRT_NT*, the intervention groups showed pre-post increases of 5.38 to 9.88 with medium within-group effect sizes (Hedge's g = 0.373 to 0.662), all *ps* <0.009. *Control* changed minimally (1.11 change). Groups did not differ significantly at *pretest* or *posttest*.

For *WSAP_NER*, the *Audio* and *Video* groups evidenced pre-post declines with medium-large within-group effect sizes (*Audio*: $MD_{(pretest-\ posttest)}$. = .211, Hedge's g = 0.713, *p* = .005; Video: $MD_{(pretest-\ posttest)}$ = .264, Hedge's g = 0.685, *p* = .001). However, the *Text* group did not show a statistically significant change across time ($MD_{(pretest-\ posttest)}$ = .203, Hedge's g = 0.657, *p* = .151). *Control* was stable (2.7 change). At *posttest*, *Audio* and *Video* differed significantly from *Control* with a medium-large between-group effect size (Hedge's g = 0.643 and 0.590, respectively).

On *WSAP_PMR*, the intervention groups showed pre-post reductions ranging from .236 to .264 with medium-large within-group effect sizes (Hedge's g = 0.604 to 0.713), all *ps* = 0.000. *Control* worsened by 20.8, though not significant. At *posttest*, *Control* remained highest and differed significantly from *Text* with a medium-large between-group effect size (Hedge's g = 0.733).

GEE analyses on the six measures were conducted again with their respective outliers excluded and found a similar effect pattern to the results with all participants (Table 5.33 and

Table 5.34). *Text* group was found to have a significant decrease on *WSAP_NER* when removed the outliers in WSAP task (E1_26, E1_35, C_4, C_11).

In summary, the interventions led to pre-post improvements across indicators with medium to large effect sizes compared to stable *Control*. Importantly, no significant baseline differences existed between groups prior to the intervention, nor the *posttest*. Therefore, the differential patterns of change from *pretest* to *posttest* demonstrate the overall effectiveness of **eiIBM_RobotV1,** regardless of **Medium**. The exclusion of the outliers strengthens the findings.

Table 5. 31 Means, Standard Errors, Percentage Change, and Effect Sizes on Intervention Outcome with N = 67 in Study 1

Medium	n	Mean (SE)		change [95 % CI]	Hedge's effect size [95 % CI]		
		pretest	posttest	pre-post	pre (versus Control)	pre-post (within)	post (versus Control)
DS_MS							
Audio	17	29.47(1.978)	14.75(3.175)	14.72 [5.94, 23.50]	0.492 [-0.042, 1.026]	1.329 [0.753, 1.904]	-0.657 [-1.236, -0.077]
Video	16	31.13(2.632)	15.69(3.110)	15.44 [10.32, 20.56]	0.827 [0.219, 1.434]	1.224 [0.596, 1.851]	-0.700 [-1.275, -0.124]
Text	16	32.50(2.592)	17.00(3.365)	15.50 [8.01, 22.99]	-1.205 [-1.896, -0.513]	1.114 [0.488, 1.739]	-0.713 [-1.254, -0.171]
Control	18	27.89(1.571)	24.17(1.698)	3.72 [-.96, 8.40]	…	…	…
SST_TNR							
Audio	17	.621(.0579)	.413(.0716)	.208[.0031, .4126]	-0.116 [-0.492, 0.260]	0.476 [0.098, 0.853]	-0.379 [-0.754, -0.004]
Video	16	.669(.0699)	.405(0655)	.264[.0784, .4501]	-0.082 [-0.457, 0.294]	0.441 [0.071, 0.810]	-0.361 [-0.736, 0.015]
Text	16	.639(.0503)	.402(.0758)	.237[.0342, .4398]	-0.023 [-0.398, 0.352]	0.414 [0.044, 0.784]	-0.403 [-0.778, -0.027]
Control	18	.628(.0520)	.565(.0531)	.063[-.0429, .1684]	…	…	…
SRT_PT							
Audio	17	23.06(1.562)	32.29(1.315)	-9.24 [-15.25, -3.22]	-0.492 [-1.026, 0.042]	-0.713 [-1.254, -0.171]	-0.710, [-1.243, -0.176]
Video	16	21.31(1.162)	27.69(1.400)	-6.37[-11.62, -1.13]	0.827 [0.219, 1.434]	0.549 [-1.018, -0.079]	-0.355 [-0.788, 0.079]
Text	16	24.19(1.596)	30.63(1.262)	-6.44[-11.70, -1.17]	0.292 [-0.242, 0.826]	-0.587 [-1.056, -0.117]	-0.604 [-1.037, -0.170]
Control	18	23.89(1.407)	23.28(1.688)	.61[-3.25, 4.48]	…	…	…
SRT_NT							
Audio	17	24.06(1.548)	14.18(1.412)	9.88[4.31, 15.46]	0.238 [-0.293, 0.768	0.662 [0.231, 1.092]	-0.482 [-0.912, -0.051]]
Video	16	21.44(1.952)	16.06(1.404)	5.38[.70, 10.05]	0.026 [-0.404, 0.456]	0.373 [-0.056, 0.802]	-0.282 [-0.712, 0.147]
Text	16	22.25(1.800)	16.00(2.278)	6.25[.88, 11.62]	0.089 [-0.351, 0.529]	0.391 [-0.038, 0.820]	-0.296 [-0.726, 0.133]
Control	18	21.11(1.157)	20.00(1.444)	1.11[-2.71, 4.93]	…	…	…
WSAP_NER							
Audio	17	.591(.0496)	.381(.0579)	.211[.0344, .3873]	-0.307 [-0.837, 0.224]	0.713 [0.171, 1.254]	-0.643 [-1.173, -0.112]
Video	16	.647(.0552)	.383(.0625)	.264[.0682, .4597]	-0.058 [-0.488, 0.372]	0.685 [0.124, 1.275]	-0.590 [-1.120, -0.059]
Text	16	.663(.0539)	.460(.0605)	.203[-.0249, .4307]	-0.011 [-0.441, 0.419]	0.657 [0.077, 1.236]	-0.433 [-0.863, -0.002]
Control	18	.661(.0339)	.634(.0471)	.027[-.0524, .1061]	…	…	…
WSAP_PMR							
Audio	17	.510(.0380)	.274(.0446)	.236[.0902, .3811]	-0.067 [-0.497, 0.363]	0.713 [0.423, 1.004]	-0.572 [-1.102, -0.041]
Video	16	.504(.0413)	.268(.0542)	.237[.0809, .3926]	-0.103 [-0.533, 0.327]	0.685 [0.318, 0.779]	-0.586 [-1.116, -0.055
Text	16	.495(.0568)	.231(.0399)	.264[.1314, .3962]	-0.162 [-0.592, 0.268]	0.604 [0.329, 0.815]	-0.733 [-1.263, -0.202]

Control	18	.525(.0242)	.440(.0369)	-.208[-.3781, -.0385]	…	…	…

Table 5. 32 Statistical Effects and Comparisons Between TestTime and Medium with N = 67 in Study 1

Indicators			Measures					
			DS_MS	SST_TNR	SRT_PT	SRT_NT	WSAP_NER	WSAP_PMR
GEE effects		TestTime	$\chi2 = 57.641$, p <.001	$\chi2 = 31.749$, p <.001	$\chi2 = 37.526$, p <.001	$\chi2 = 46.176$, p <.001	$\chi2 = 30.495$, p <.001	$\chi2 = 62.820$, p <.001
		Medium	$\chi2 = 3.331$, p =.343	$\chi2 = 2.066$, p =.559	$\chi2 = 7.784$, p =.051	$\chi2 = 1.950$, p =.581	$\chi2 = 9.320$, p =.025	$\chi2 = 10.453$, p =.015
		TestTime × Medium	$\chi2 = 24.640$, p <.001	$\chi2 = 12.628$, p =.006	$\chi2 = 21.437$, p <.001	$\chi2 = 18.465$, p <.001	$\chi2 = 19.842$, p <.001	$\chi2 = 21.265$, p <.001
p value from pairwise comparisons	pre (between)	Audio versus Video	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Text versus Control	1.000	1.000	1.000	1.000	1.000	1.000
	pre-post	Audio pre -> post	.000	.042	.000	.000	.005	.000
		Video pre-> post	.000	.000	.004	.009	.001	.000
		Text pre -> post	.000	.007	.004	.008	.151	.000
		Control pre -> post	.364	1.000	1.000	1.000	1.000	.222
	post (between)	Audio versus Video	1.000	1.000	.460	1.000	1.000	1.000
		Audio versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	.250	1.000	.001	.110	.020	.120
		Video versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Control	.468	1.000	1.000	1.000	.038	.243
		Text versus Control	1.000	1.000	.014	1.000	.659	.004

Table 5. 33 Means, Standard Errors, Percentage Change, and Effect Sizes on Intervention Outcome with Outliers Excluded in Study 1

Medium	n	Mean (SE)		change [95 % CI]	Hedge's effect size [95 % CI]		
		pre	post	pre-post	pre (versus *Control*)	pre-post (within)	post (versus *Control*)
DS_MS							
Audio	17	29.47 (1.978)	14.51 (3.169)	14.96 [6.40, 23.53]	0.705 [-0.274, 0.519]	1.341 [0.765, 1.916]	-0.659 [-1.238, -0.079]
Video	16		15.69 (3.110)	15.44 [10.32, 20.56]	0.827 [0.219, 1.434]	1.225 [0.597, 1.852]	-0.702 [-1.277, -0.126]
Text	16	31.12 (2.632)	17.00 (3.365)	15.50 [8.01, 22.99]	-1.206 [-1.897, -0.514]	1.116 [0.489, 1.742]	-0.713 [-1.254, -0.171]
Control	17		24.47 (1.771)	2.65 [-0.91, 6.20]	…	…	…

		32.50 (2.592)					
		27.12 (1.462)					
SST_TNR							
Audio	14	.629 (.0621)	.401 (.0770)	.229 [.0198, .4373]	0.002 [-0.468, 0.472]	0.505 [0.026, 0.983]	-0.362 [-0.741, 0.016]
Video	15	.669 (.0699)	.435 (.0600)	.234 [.1024, 3649]	-0.082 [-0.457, 0.294]	0.441 [0.071, 0.810]	-0.252 [-0.631, 0.127]
Text	16	.639 (.0503)	.397 (.0752)	.242 [.0405, .4429]	-0.023 [-0.398, 0.352]	0.414 [0.044, 0.784]	-0.357 [-0.736, 0.021]
Control	18	.628 (.0521)	.565 (.0531)	.063 [-.0429, .1684]	…	…	…
SRT_PT		23.87 (1.638)				-0.713 [-1.254, -0.171]	-0.710 [-1.241, -0.178]
Audio	15	21.87 (1.099)	32.20 (1.440)	-8.33 [-14.47, -2.19]	0.066 [-0.365, 0.497]	-0.570 [-1.043, -0.096]	-0.378 [-0.809, 0.053]
Video	15	24.93 (1.725)	28.27 (1.368)	-6.40 [-11.99, -.81]	0.888 [0.357, 1.418]	-0.612 [-1.086, -0.137]	-0.568 [-0.999, -0.136]
Text	14	23.76 (1.485)	31.21 (1.371)	-6.29 [-12.28, -.30]	0.637 [0.006, 1.267]	…	…
Control	17		23.65 (1.747)	.12 [-.366, 3.89]	…		
SRT_NT		22.80 (1.415)				0.680 [0.250, 1.110]	-0.366 [-0.796, 0.065]
Audio	15	20.73 (1.952)	14.47 (1.514)	8.33 [4.00, 12.67]	0.258 [-0.272, 0.789	0.341 [-0.088, 0.770]	-0.324 [-0.754, 0.107]
Video	15	22.64 (2.029)	15.53 (1.395)	5.20 [.24, 10.16]	0.050 [-0.381, 0.481]	0.411 [-0.019, 0.841]	-0.283 [-0.713, 0.148]
Text	14	21.11(1.157)	15.57 (2.581)	7.07 [1.30, 12.84]	0.231 [-0.300, 0.762]	…	…
Control	17		19.94 (1.528)	.71 [-3.12, 4.53]	…		
WSAP_NER		.5914 (.0496)				0.713 [0.171, 1.254]	-0.638 [-1.068, -0.207]
Audio	17	.6531 (.0585)	.3806 (.0579)	.2108 [.0344, .3873]	-0.313 [-0.743, 0.116]	0.642 [0.112, 1.172]	-0.636 [-1.066, -0.205]
Video	15	.6937 (.0478)	.3604 (.0603)	.2927 [.1152, 4703]	-0.036 [-0.466, 0.394]	0.723 [0.293, 1.153]	-0.372 [-0.802, 0.058]
Text	15	.6586 (.0348)	.4470 (.0631)	.2467 [.0489, .4444]	0.132 [-0.298, 0.562]	…	…
Control	16		.6128 (.0485)	.0456 [-.0317, .1232]	…		
WSAP_PMR		.5100 (.0381)				0.713 [0.171, 1.254]	-0.576 [-1.106, -0.045]
Audio	17	.5025 (.0440)	.2743 (.0446)	.2357 [.0902, .3811]	-0.081 [-0.511, 0.349	0.686 [0.116, 1.255]	-0.570 [-1.100, -0.039]
Video	15	.4776 (.0578)	.2785 (.0564)	.2240 [.0621, .3859]	-0.106 [-0.536, 0.324]	0.628 [0.058, 1.197]	-0.750 [-1.280, -0.219]
Text	15	.5181 (.0264)	.2253 (.0421)	.2523 [.1195, .3851]	-0.170 [-0.600, 0.260]	…	…
Control	16		.4333 (.0410)	.0848 [-.0251, .1948]	…		

Table 5. 34 Statistical Effects and Comparisons Between TestTime and Medium with Outliers Excluded in Study 1

Indicators			Measures					
			DS	SST_TNR	SRT_PT	SRT_NT	WSAP_NER	WSAP_PMR
GEE effects		TestTime	$\chi2 = 57.090$, $p < .001$	$\chi2 = 34.531$, $p < .001$	$\chi2 = 32.678$, $p < .001$	$\chi2 = 40.465$, $p < .001$	$\chi2 = 41.659$, $p < .001$	$\chi2 = 57.516$, $p < .001$
		Medium	$\chi2 = 3.060$, $p = .382$	$\chi2 = 1.943$, $p = .584$	$\chi2 = 8.321$, $p = .040$	$\chi2 = 2.098$, $p = .552$	$\chi2 = 8.072$, $p = .045$	$\chi2 = 8.644$, $p = .034$
		TestTime × Medium	$\chi2 = 32.739$, $p < .001$	$\chi2 = 14.507$, $p = .002$	$\chi2 = 16.559$, $p < .001$	$\chi2 = 18.728$, $p < .001$	$\chi2 = 22.724$, $p < .001$	$\chi2 = 17.357$, $p < .001$
p value from pairwise comparisons	pre (between)	Audio versus Video	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Text	1.000	1.000	1.000	1.000	1.000	1.000

		Audio versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Text versus Control	1.000	1.000	1.000	1.000	1.000	1.000
	pre-post	Audio pre -> post	.000	.017	.001	.000	.005	.000
		Video pre-> post	,000	.000	.010	.029	.000	.000
		Text pre -> post	.000	.005	.029	.004	.003	.000
		Control pre -> post	.559	1.000	1.000	1.000	1.000	.447
	post (between)	Audio versus Video	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	.170	1.000	.004	.306	.059	.244
		Video versus Text	1.000	1.000	1.000	1.000	1.000	1.000
		Video versus Control	.395	1.000	1.000	.928	.031	.741
		Text versus Control	1.000	1.000	.018	1.000	1.000	.011

5.3.8 Effects of Medium on Intervention Outcome Differences Over Time

Previous GEE analyses showed that the intervention groups demonstrated significant improvements after the intervention phase. In this section, I further analyzed whether there were differences in intervention effects caused by **Medium** using residual change scores of the experiential variables.

The residual change scores obtained were *RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, *RES_NT* and *RES_PT* for *DS_MS*, *SST_TNR*, *WSAP_NER*, *WSAP_PMR*, *SRT_NT*, and *SRT_PT* between pretest and posttest, respectively. For *RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, and *RES_NT*, a residual change score greater than 0 indicates the *posttest* score was higher than predicted, meaning it did not decrease to the expected level after intervention. The larger the value, the farther from the expected value after intervention. In contrast, for *RES_PT,* a residual change score greater than 0 indicates positive change from *pretest* to *posttest*.

Test of normality on the residual change scores across **Medium** showed most data were normally distributed, except for *RES_PT* and *RES_PMR* in the *Text* group and

RES_TNR in the *Video* group. Thus, one-way MANOVA analyses on the residual change scores of different intervention groups (*Audio*, *Video* and *Text*) were conducted and the results showed the composite dependent variable was not significantly affected by **Medium**, F (12, 84) = 1.079, p = .388, Pillai's V = .267, partial $\eta 2$ = .134. Follow-up one-way ANOVAs indicated no significant differences between **Medium** on all measures (all ps > .05). These results suggest participants benefited from the **eiIBM_RobotV1** in symptoms of depression and negative interpretations biases, regardless of robot types. The results remained the same when experiential outliers were removed.

5.3.9 Effect of Experience on Intervention Outcome Difference

Next, I aimed to understand the effect of overall experience on the intervention outcomes. Due to the small size (N =49), it is impractical to run the regression analyses with all experience measures over three times as predictors. Instead, hierarchical cluster analysis (HCA) was used to identify participant groups with homogeneous experience patterns (Zhang et al., 2017).

Ward's method with squared Euclidean distance was used to categorize participants based on experience ratings across 24 measures over three times. A 3-cluster solution was found optimal: 'high experience' (*Exp_H*, n =16), '*medium experience*' (*Exp_M*, n =12), and "*low experience*" (*Exp_L*, n =21). The experience clusters formed a new ranking variable *ExpRank*. The *Control* group was excluded as they did not have experience data.

Almost all residual depressive measures change scores were normally distributed within their *ExpRank* (except *Exp_L*'s RES_DS: p = .043). Thus, a one-way MANOVA assessed *ExpRank* differences in intervention outcome. The composite dependent variable was significantly affected by *ExpRank*, F (48, 48) = 2.55, p < .001, Pillai's V = 1.436, partial $\eta 2$ = .718. Follow-up one-way ANOVAs showed significant differences between *ExpRank* on

all experience measures (all *p*s <.001), with the *Exp_H* cluster having the highest average ratings, followed by *Exp_M* cluster and the *Exp_L* cluster, all *ps* <.05 (Table 5.35). Pairwise comparison indicated most of the difference (60 out 72 pairwise comparison) were significant.

Table 5. 35 Means on Experience Measure across ExpRank in Study 1

	T1			***T2***			***T3***		
Measures	Exp_L	Exp_M	Exp_H	Exp_L	Exp_M	Exp_H	Exp_L	Exp_M	Exp_H
M_AffEase	4.60 (.98)	5.10 (.60)	5.59 (.55)	4.58 (.75)	4.92 (.68)	5.69 (.42)	4.23 (.87)	4.92 (.79)	5.67 (.43)
M_RelEase	3.94 (.70)	4.67 (.67)	5.41 (.58)	4.07 (.76)	4.67 (.58)	5.44 (.62)	4.00 (.88)	4.63 (.75)	5.48 (.51)
M_ValEase	3.95 (.70)	4.47 (.72)	5.33 (.54)	3.92 (.70)	4.64 (.61)	5.54 (.42)	3.75 (.68)	4.92 (.29)	5.40 (.53)
M_ValEmo	3.60 (.69)	4.13 (.56)	5.00 (.66)	3.73 (.75)	4.23 (.74)	5.25 (.43)	3.63 (.68)	4.50 (.46)	5.42 (.42)
M_UseIntP	3.76 (.76)	4.19 (.39)	4.85 (.52)	3.73 (.81)	3.97 (.39)	5.31 (.54)	3.5 (.87)	4.44 (.61)	5.17 (.60)
M_TrustP	3.76 (.78)	4.25 (.58)	5.25 (.55)	3.81 (.75)	4.37 (.64)	5.50 (.48)	3.71 (.75)	4.42 (.63)	5.41 (.49)
M_UseP_i	4.02 (.75)	4.75 (.40)	5.31 (.60)	4.00 (.77)	4.79 (.58)	5.41 (.58)	3.62 (.93)	4.71 (.45)	5.56 (.48)
M_UseP_ci	3.81 (.97)	4.67 (.54)	5.44 (.57)	3.90 (.92)	4.79 (.45)	5.69 (.40)	3.90 (.87)	4.79 (.50)	5.78 (.36)
Num	21	12	16	21	12	16	21	12	16

Note. Mean (SD) in each cell.

Using *ExpRank* categorization, another one-way MANOVA checked differences in intervention effect using residual change scores. The composite intervention effect significantly differed between *ExpRank* groups, F (12, 84) = 2.169, p = .021; Pillai's V = .473, partial $\eta 2$ = .237. Follow-up ANOVAs showed significant differences on *RES_TNR*, *RES_NER*, *RES_NT*, and *RES_PT*, but not *RES_PMR* or *RES_DS* (Table 5.36).

Table 5. 36 Univariate Test of ExpRank on Experiential Residual Change with N = 49 in Study 1

Effect	*F*	Hypothesis *df*	Error *df*	*Sig.*	Partial Eta Squared
RES_TNR	3.827	2	46	.029	.143
RES_NER	7.151	2	46	.002	.055.237
RES_NT	6.608	2	46	.003	.223
RES_PT	3.484	2	46	.039	.132
RES_PMR	2.571	2	46	.087	.101
RES_DS	2.503	2	46	.093	.098

Post-hoc Bonferroni comparisons (Table 5.37) indicated that for *RES_TNR*, *Exp_H* (M = -.083, *SD*. = 0.223) group showed greater decreases to the *Exp_L* group (M = .105, *SD*

= .228), $MD_{(Exp\_H- Exp\_L)}$ = -0.189, p = 0.053, Cohen's d = -2.455. On *RES_NT*, the *Exp_H* group decreased more compared to the *Exp_M* group ($MD_{(Exp\_H- Exp\_M)}$ = -6.360, p = 0.012, Cohen's d = -3.252) and the *Exp_L* group ($MD_{(Exp\_H- Exp\_L)}$ = -5.958, p = 0.006, Cohen's d = -3.253). For *RES_NER*, the *Exp_H* group decreased significantly more than the *Exp_L* group ($MD_{(Exp\_H- Exp\_L)}$ = -0.263, p = 0.002, Cohen's d = -3.757). On *RES_PT*, the *Exp_H* group improved more than the *Exp_L* group ($MD_{(Exp\_H- Exp\_L)}$ = 4.506, p = .043, Cohen's d = 2.544). However, the effect sizes for these differences were small. These results suggest higher experience levels relate to greater residual decreases in negative interpretation biases and increase in positive target endorsement after **eiIBM_RobotV1**. Comparable results were found when removing therapy effects outliers.

Table 5. 37 Pairwise Comparison between ExpRank on Residual Change of Pre-Post Assessment with N = 49 in Study 1

Measure	I-ExpRank	J-ExpRank	*MD(I-J)*	St. Error	Cohen's d	Sig.*
RES_DS	*Exp_H*	*Exp_M*	-5.192	3.363	-1.544	0.106
	Exp_H	*Exp_L*	-6.341	2.922	-2.170	0.388
	Exp_M	*Exp_L*	-1.149	3.19	-0.360	1.000
RES _TNR	*Exp_H*	*Exp_M*	-0.020	0.088	-0.227	1.000
	Exp_H	*Exp_L*	-0.189	0.077	-2.455	0.053
	Exp_M	*Exp_L*	-0.179	0.084	-2.131	0.114
RES _PT	*Exp_H*	*Exp_M*	3.803	2.038	1.866	0.205
	Exp_H	*Exp_L*	4.506	1.771	2.544	0.043
	Exp_M	*Exp_L*	0.703	1.931	0.364	1.000
RES _NT	*Exp_H*	*Exp_M*	-6.360	2.108	-3.017	0.012
	Exp_H	*Exp_L*	-5.958	1.832	-3.252	0.006
	Exp_M	*Exp_L*	0.401	1.998	0.201	1.000
RES _NER	*Exp_H*	*Exp_M*	-0.111	0.081	-1.370	0.534
	Exp_H	*Exp_L*	-0.263	0.070	-3.757	0.002
	Exp_M	*Exp_L*	-0.153	0.077	-1.987	0.158
RES _PMR	*Exp_H*	*Exp_M*	-0.094	0.061	-1.541	0.387
	Exp_H	*Exp_L*	-0.116	0.053	-2.187	0.097
	Exp_M	*Exp_L*	-0.023	0.058	-0.397	1.000

Note. Adjusted by the Bonferroni correction.

5.5 Summary and Discussion

This study utilized three types of robots (audio robot, telepresence robot, and chatbot) to deliver an elaborative interpretation bias modification (**eiIBM**) intervention, **eiIBM_RobotV1**. It explored user's experiential differences with varied robot agents and examined their impacts on intervention outcomes. The I-PEFiC guided assessments of experiential variables. Intervention outcomes were evaluated using depressive symptoms scales (BDI-II) and interpretations screening tasks assessing automatic (WSAP-D), elaborative (SRT), and medium-speed (SST) processing.

The study addressed three research questions:

RQ1.1: How do user perceptions differ between interacting with different types of robots (dissimilarity hypothesis) or are the perceptions largely similar (similarity hypothesis)?

RQ1.2: Given **eiIBM_RobotV1**, do the intervention outcomes differ between interactions with different robot types?

RQ1.3: How do user perceptions influence intervention outcomes?

5.5.1 Discussion on Research Questions

Thorough analyses tested the hypotheses. Regarding experiential variables, participants' perceptions were highly consistent across **eiIBM_RobotV1** delivered by different robots in terms of intervention delivery affordance and ease-of-use interaction affordance, except slightly higher initial use intention for audio and telepresence robots versus chatbot. This supported similarity hypotheses **H5.2** while rejecting the dissimilarity hypotheses **H5.3**-**H5.6**. All intervention groups significantly reduced depressive symptoms and negative interpretation biases compared to the control group, validating **H5.7**. More positive experiences overall associated with greater reduction in negative interpretation biases and depressive symptoms, validating **H5.8**.

In summary, results evidenced highly similar user experiences across **eiIBM_RobotV1** delivered by different robot agents (**RQ1.1**). With descriptive statistics, While the chatbot started with lower experiential ratings, these gradually increased over time, indicating growing acceptance. The telepresence robot afforded a better initial experience that declined, while the audio robot remained relatively stable. However, GEEs and Bayesian analyses revealed no significant experiential differences across intervention groups or timepoints overall, indicating highly consistent user experiences. Slightly higher initial use intention for audio-guided and video-guided **eiIBM_RobotV1** could be due to movement and appearance cues enhancing attraction (Almeida et al., 2022), which normalized over time. Measurement invariance analyses demonstrated no marked divergences in experiential evaluation paths across timepoints.

No significant intervention outcome differences emerged across **eiIBM_RobotV1** guided by different robot agents (**RQ1.2**). GEEs on intervention outcome found similar pre-post change trends and residual score changes, indicating no intervention effects caused by the robots delivering the **eiIBM_RobotV1**. This suggests comparable robot impacts on the depression and negative bias interpretation bias reduction given equivalent intervention mechanism. Intervention outcomes depended on mechanism rather than medium.

As expected, more positive experiences were associated with reduced negative biases (**RQ1.3**). Cluster analysis showed that the high experience group had the greatest negative bias decrease, indicating experience facilitated effects. Key indicators were reduced medium-speed processing negativity bias (*SST_TNR*), increased elaborative positive target endorsement (*SRT_PT*), and decreased automatic negative bias endorsement (*WSAP_NER*). All suggested that residual changes correlated with experience level.

5.5.2 Theoretical and Practical Implications

Theoretically, this study validated applying the I-PEFiC to understand robot-delivered intervention experiences. It underscored the importance of addressing primary user goals through affordances, aligning with Hoorn and Huang (2024). I-PEFiC holds that varied robot feature encoding elicits different affordances, comparisons, and recursive responses influencing subsequent perceptions. Path analysis demonstrated that intervention delivery affordance strongly influenced use intention and trust when the emotional distress relief was highly possible (among depressed participants), while ease-of-use interaction affordances were less impactful. This evidenced the significance of task-contingency. PCA indicated experiential evaluations evolve over time with growing understanding, emphasizing the need for repeated measurements in human-robot experiments.

Practically, the research provided an effective online **eiIBM_Robot** intervention allowing flexible robot modalities choices. Although not prominent, GEEs and Bayesian analyses revealed increased emotional distress relief valence over time across robots, suggesting potentially growing benefit awareness with sustained use intention. Audio and telepresence robots might mitigate early dropout, which is critical for persisting to detect utility in real-world applications. Thus, although audio and telepresence did not improve outcomes or overall experience over time, they could facilitate participant retention until the intervention takes effect.

Another practical implication for design is that Task-contingency design can contribute to the overall experience benefiting intervention outcomes. This study revealed higher intervention delivery affordance valence scores led to higher use intention, trust, usefulness evaluations, and thus overall experience. Analyses of experience rank effects on intervention outcomes showed higher overall experience gains related to improved

intervention outcomes. Results align with another empirical evidence that although voice interactions may feel more enjoyable, text interactions achieve better outcomes (Bickmore et al., 2018; Vaidyam et al., 2019). Added robot modalities can provide interesting interactions but do not necessarily improve overall engagement and use intention. However, engagement and use are pivotal for outcomes. This suggests task-contingency design optimally serving user goals, rather than assuming more modalities give better experiences and outcomes, can improve both overall experience and intervention outcomes.

5.5.3 Limitations and Future Research

The current study faced some limitations that should be addressed in future research. The small sampling size precluded path analysis by robot modalities. Also, the low statistical power increased the likelihood of failing to detect true effects (Type II error), though this the maximum recruitable given limited time.

The author had originally planned to perform a Repeated Measures MANOVA and conducted a priori power analysis for this approach. However, during the actual data analysis, it was found that the data did not meet the assumptions for MANOVA, such as normality. As a result, the analysis plan was pivoted to use Generalized Estimating Equations (GEE), which are more robust to deviations from normality.

The post-hoc power analysis conducted using the GEE models revealed that the statistical power ranged from 0.03 to 0.69 across the different outcome measures. This lower than desired power is likely due to the small cluster sizes and the lack of reliable prior data to inform the parameter estimates used in the power calculations.

Further research are welcome to replicate with larger samples to enhance generalizability and enable more granular analyses. The lack of a placebo group meant improved self-reported symptoms could potentially emerge from placebo effects. However,

the rapid forced choices required in WSAP-D and SST likely minimized deliberate positive bias, as depression manifests in automatic rather than attitudinal processing due to depressive schemas. Self-report provided the sole experiential measure, potentially constraining score variations due to moderacy and central tendency bias (Saris & Gallhofer, 2007) or individual difference noise stemming from specific robots' communication cues and counteracting facet experiences.

In path analysis, ease-of-use interaction valence significantly predicted use intention and trust at *T1* and *T2*, while relevance did not, suggesting perceived goal achievement capability mattered more than relevance. However, the measures on ease-of-use interaction valence only captured post-disconfirmation / confirmation results lacking initial expectation reactions. This precluded conclusions about depressed participants various strategies for addressing expectation confirmation.

Follow-up research in Chapter 6 will address these limitations through a within-subject experiment (audio, telepresence, and chatbot robots) and semi-structured interviews to optimize understanding. The interviews will elicit decision processing details on **eiIBM_RobotV1** use intention and trust to confirm the I-PEFiC path and shed light on improving valence and use intention/trust by examining expectation confirmation reactions. This approach will provide greater insight into individual differences in perceptions and outcomes across robot modalities, contributing to the incorporation of robots into the standardized therapy.

Chapter 6: Experience of eiIBM_RobotV1 Guided by Different Robots (Study 2)

In this chapter, I studied the underlying reasons for the experience and intervention outcome differences or similarities across robot modalities by exploring the individuals' experience on robots guiding them the **eiIBM_RobotV1**.

6.1 Introduction

Study 1 (Chapter 5) found statistical similarities in experiential variables across **eiIBM_RobotV1** delivered by three diverse types of robots: audio bot, telepresence robot and chatbot. The similar experiences elicited by the audio bot and telepresence robot were unsurprising, as they included minimal but valid features that composite affordances relevant to depressed individuals' goals and positively facilitated their expectation of achieving those goals. However, the reasons behind people experiencing the chatbot similarly to the telepresence robot and audio bot remain to be explored, since it was assumed that the chatbot lacks features enabling ease-of-use interaction.

Path analyses results revealed that valence contributed more to use intention and trust in the program compared to the relevance of the affordances and goals. As valence was measured as the end-state of comparing affordance to the goal, it could not provide details into how depressed participants compare and deal with expectation conformity/disconformity.

The current within-subject study, followed by a semi-structured interview, has two research aims. The first was to evaluate the similarity of experiences when an individual interacts with different robots, improving Study 1's credibility in case of distorted findings without sufficient sample size. The second was to explore how participants compare affordances to goals regarding valence and relevance, aiming to understand the reasons

behind similar experiences with different robots. The following research questions were formulated:

RQ2.1: Does similarity of experience occur within an individual interacting with different robots?

RQ2.2: How do depressed individuals compare affordances to goals and concerns in terms of valence and relevance?

The experimental instruments and settings were identical to Study 1, with the difference being the study design. In Study 1, individuals were randomly assigned to one **Medium** condition and interacted with an assigned robot several times. In contrast, in Study 2, each individual interacted with three mediums in separate sessions over one week, with 2–3-day intervals between sessions. An additional component in Study 2 was participants being interviewed about comparing the three **Mediums** after all interactions. Their perceptions and comparisons across interactions could provide insight into their preferences and considerations regarding the robots guiding **eiIBM_RobotV1**, as well as insights into how to improve the program, thus answering **RQ 2.1** and **RQ 2.2**.

6.2 Methods

6.2.1 Participant

The inclusion criteria for participants in Study 2 were identical to those in Study 1. Participants were recruited between March 2023 and April 2023 using the same methods as in Study 1. The participants in Study 2 did not join Study 1 previously. Eligible participants were asked to sign the consent form after reading the information sheet that briefed them on the study details (e.g., aims, length, involvement, randomization, incentives). Finally, 40 depressed residents of Hong Kong (M_{age}= 24.73, SD_{age}= 3.234, 27 Female, 35 Cantonese

speaker) completed the experiments and received HK$150 ParkShop cash coupons as compensation.

6.2.2 Procedure

The study utilized a within-subject design with **Medium** (*Video* / *Audio* / *Text*) as the independent variable. Participants interacted with three mediums for over one week, with 2–3-day breaks between sessions. In each session, participants interacted with one of the robots through the assigned medium exposure sequence, experiencing an interaction procedure similar to Study 1. After each interaction, they completed an experience questionnaire about the specific robot. The questionnaire had to be finished within 10 minutes of ending the interaction to be deemed valid.

Unlike Study 1, participants did not need to complete pre-post cognitive assessments. Instead, they participated in a 10 – 30 minutes interview after finishing all interactions. The experience of interacting with a previous robot type and the order of exposure might influence participants' evaluation of a specific robot (c.f. Bradley, 1958; Morii et al., 2017). To minimized potential anchoring effects or order effects from the sequence of robot exposure (Charness et al., 2012), the order of mediums was randomized across participants using six possible sequences: *Audio-Video-Text*, *Audio-Text-Video*, *Video-Text-Audio*, *Video-Audio-Text*, *Text-Audio-Video*, *Text-Video-Audio*.

6.2.3 Apparatus and Materials

Study 2 utilized the **eiIBM_RobotV1** developed in Study 1, along with its task stimulus, interactive protocol, and operational settings. However, only the stimulus from sessions 1, 3 and 5 were used. The three robots were named "XiaoCong" (audio bot), "XiaoZhi" (telepresence robot), and "XiaoWei" (chatbot) to distinguish them. Before each exercise, the robots introduced themselves and guided participants to complete the **eiIBM**

exercise in their respective capable ways. In the first session, the robot also led an imagination exercise. For example, a participant assigned to *Audio-Video-Text* exposure order would have the audio bot to teach him/her using imagination in the following exercise.

6.2.4 Measures

The measures focused on participants' experience with the **Medium** to align with Study 2's aims. In Study 1, there were two key affordances – intervention delivery (*AffEmo*) and ease-of-use interaction (*AffEase*). *AffEmo* was derived on **eiIBM** exercise format (*eiIBM_Implementation*) and robot **(***eiIBM_Medium*) features jointly, while *AffEase* majorly derived from robot features specifically. Questions regarding valence/relevance of *AffEmo* targeted the program overall (e.g., the subject of the question statement was "the program"), while questions regarding *AffEase* and its valence/relevance targeted the specific robots (e.g., the subject of the question statement was "XiaoZhi"). Study 1 found affordances to emotional relief was most important for use intentions and trust on the overall program.

Study 2, therefore, focuses on ease-of-use affordance (from *eiIBM_Medium*) and the use intention to specific robots (from *eiIBM_Medium*), while further exploring the correlation between robot use intention and the evaluation on the **eiIBM_RobotV1** as a whole. The following variables were measured: ease-of-use affordance (*AffEase*), relevance and valence of ease-of-use for an enjoyable experience (*RelEase* and *ValEase*), intention to use the robot again (*UseIntR*), usefulness evaluation (*UseP*) and trust (*TrustP*) in **eiIBM_RobotV1**. In summary, six scales were measured: *AffEase*, *ValEase*, *RelEase*, *UseIntR*, *UseP* and *Trust*, using 6-point Likert scales (strongly disagree = 1, strongly agree = 6). Items were presented randomly within blocks. Table 6.1 displays the scale descriptions. *Gender*, *Age*, *Language* were collected as control factors during registration.

Five 10-point rating scales measured manipulations checks on the ease (*Check_EaseImg*) and engagement (*Check_EngImg*) for the mental imagery, overall exercise engagement (*Check_EngExcer*), robot satisfaction (*Check_SatR*) and meeting expectations (*Check_ExpR*).

Table 6. 1 Items for Scale Measuring Experiential Variables in Study 2

Construct	Categories	Experience source	
Affordance	Affordance – Ease-of-use (*AffEase*)	Robot	*AffEase#_1*: Based on my experiences, the training program's interactions are clear and easy to understand *AffEase#_2*: Based on my experiences, I find that such training programs are easy to use *AffEase#_3*: Based on my experiences just now, I find that I can easily become proficient in the interaction process *AffEase#_4R*: Based on my experience just now, - it would take me a long time to get used to such a training program (R) *AffEase#_5*: Based on my experiences, I immediately understand how I should interact with the training program *AffEase#_6R*: Based on my experiences, the training program is a little difficult to use (R)
Relevance	Relevance – robot relevant to the goal of completing the exercise (*RelEase*)	Robot	*RelEase#_1*: In terms of providing help, robot - makes me complete tasks faster *RelEase#*_2: In terms of providing help, robot - improves my efficiency *RelEase#_3*: In terms of providing help, robot - helped me get involved in the training *RelEase#_4*: In terms of providing help, robot - enhances the training effect *RelEase#_5R*: In terms of providing help, robot - increased the difficulty of my training (R)
Valence	Valence – expectation on the comfortable experience (*ValEase*)	Robot	*ValEase#*_1: The experience with [name of robot] (guide), I felt – it is pleasant *ValEase#_2*: The experience with [name of robot] (guide), I felt - it is interesting *ValEase#_3R*: The experience with [name of robot] (the guide), I felt - is uninteresting (R) *ValEase#_4R*: The experience with [name of robot] (the guide), I felt - it is disappointed (R) *ValEase#_5*: In the experience with [name of robot] (guide), I felt – is happy
Use Intention	Use intention – robot (*UseIntR*)	Robot	*UseIntR#_1*: Based on my experience just now, I am willing to have this robot to help me *UseIntR#_2*: Based on what I just experienced - I plan to continue have this robot to help me *UseIntR#_3*: Based on what I just experienced - I would like to have this robot to help me *UseIntR#_4*: Based on what I have just experienced – I am open to have this robot to help me

			UseIntR#_5: Based on what I have just experienced – I will consider having this robot to help me
Trust	Engagement in the exercise (*TrustP*)	Robot and exercise format	*TrustP#_1*: The following is a rating of trust in the training program. Overall, I think it's reliable *TrustP#_2*: The following is a rating of trust in the training program. On the whole, I can count on it *TrustP#_3*: The following is a rating of trust in this training program. Overall, it is capable *TrustP#_4*: The following is a rating of trust in the training program. Overall, I - am confident that it can helps me *TrustP#_5R*: The following is a rating of trust in the training program. Overall, it is unreliable (R)
Satisfaction	Usefulness of the program (*UseP*)	Robot and exercise format	*UseP#_1*: I think such a training program is useful *UseP#_2R*: I think such training program is not valuable(R) *UseP#_3*: I think such training programs is good for me *UseP#_4R*: I think this training program change nothing (R) *UseP#_5*: I think this training program is beneficial
Manipulation check			Check_EngImg#: The extent to which I actively imagined the scene described by [name of robot] in the process was Check_EaseImg#: The degree to which I have no trouble imagining these scenarios is: Check_EngExcer#: In general, the degree to which I actively participate in this training is as follows: Check_SatR#: In general, my satisfaction with [name of robot]'s performance is as follows: Check_ExpR#: As an audio/text/telepresence robot, I think [name of robot] lived up to my expectations of it

Note. # in variable name represent Medium: T → Text, A → Audio, V → Video

6.2.5 Interviews

Semi-structured interviews were conducted in person at the participant's chosen location or by phone for those reporting anxiety in public settings. All interviews were audio recorded, transcribed for analysis, and lasted 10 – 30 minutes, averaging 20 minutes.

Open-ended questions explored participants' occupation, emotional distress experiences, attempted coping strategies, study involvement, and perspectives on robot-assisted therapy applications. This established contextual background regarding participants' roles and emotional distress types.

Subsequent questions examined experiences across robot mediums, preferences, perceived differences, willingness for future online training, and criteria for trusting exercise robots. While following a natural flow, the interviewer ensured all questions were covered. The following topics were investigated:

1. Whether emotional distress types and coping strategies influence robot preferences and overall experience? (Q2 and Q3)
2. Whether understanding of and alignment with study purposes and session goals impacts engagement level? (Q4, Q6 and Q7)
3. How do experiences differ across robot mediums (chatbot, audio, telepresence) for **eiIBM_RobotV1**? (Q8, Q9 and Q10)
4. Whether mismatches between personal experiences and positively resolved exercise scenarios causes discomfort? (Q11)
5. What are preferences for future online training - robot assistance or human guidance? (Q12)
6. How is the trustworthiness of exercise robots perceived and what factors are considered in assessing trust? (Q13)

6.3 Data Analysis Plan and Preparation

6.3 1 Quantitative Data

6.3.1.1 Reliability Analyses

Before conducting reliability test, the counter-indicative items were converted (1→6, 2→5, 3→4, 4→3, 5→2, 6→1) and appended with "*R*". Session 1data was used for reliability and validity analyses, assuming participants' understanding of the scales remained consistent across sessions. Items were suffixed with "*_S1*" to denote Session 1 measures. Pre-check for acquaintance bias identified participants #66, #146, #150 and #162 with inconsistent

responses to counter-indicator items. They were excluded from reliability analysis to avoid bias ($N_{(S1)}$ = 36). Reliability analysis was performed iteratively, removing items until Cronbach's $\alpha > .79$ (Nunnally, 1975) for scales with a minimum scale length of 3 or Spearman-Brown coefficient larger than 0.6 (Ursachi et al., 2015) for 2-item scales. able 6.2 lists the Cronbach's α for the shortened scales.

Table 6. 2 Reliability of Scales for Session 1 in Study 2

Scale	Items	Items	Alpha / *r*	Standardized alpha	Item means	Item variances	*N*
AffEase_S1	*AffEase_1_S1, AffEase_2_S1, AffEase_3_S1, AffEase_6R_S1*	3	.84	.84	4.94	.66	36
RelEase_S1	*RelEase_3_S1, RelEase_4_S1*	2	*r* = .88 (<.001)				36
ValEase_S1	*ValEase_1_S1, ValEase_2_S1, ValEase_3R_S1, ValEase_4R_S1, ValEase_5_S1*	5	.93	.93	4.16	1.79	36
UseIntR_S1	*UseIntR_1_S1, UseIntR_2_S1, UseIntR_3_S1, UseIntR_5_S1*	4	.97	.97	4.08	1.49	36
UseP_S1	*UseP_1_S1, UseP_2R_S1, UseP_3_S1, UseP_4R_S1, UseP_5_S1*	5	.94	.94	3.82	1.54	36
TrustP_S1	*TrustP_1_S1, TrustP_2_S1, Trust_3_S1, Trust_4_S1*	4	.85	.85	4.23	1.06	36

6.3.1.2 Validity Analyses - Principal Component Analysis (PCA)

PCA with Promax rotation and factor loadings < .50 suppressed was conducted on the remaining items to check factor structure. The PCA enabled extraction of the underlying dimensions within the session 1 variables. KMO (=.700) and Bartlett's Test (Approx. Chi-square = 964.53) indicated adequate sample quality for PCA, with 4 components explaining 77.76% variance. However, *TrustP_S1* items cross-loaded and were removed to improve validity and purify measures. Without *TrustP_S1* item, the resulting PCA had KMO = .77 and 80.06% explained variance (Table 6.3).

Component 1 had strong loadings from 5 usefulness items (*UseP_1_S1*, *UseP_2R_S1*, *UseP_3_S1*, *UseP_4R_S1*, *UseP_5_S1*) and 4 robots use intention items (*UseIntR_1_S1*, *UseIntR_2_S1*, *UseIntR_3_S1*, *UseIntR_5_S1*), indicating a high correlation between use intention and perceived usefulness of **eiIBM_RobotV1**. Component 2 contained 5 valence

items (*ValEase_1_S1*, *ValEase_2_S1*, *ValEase_3R_S1*, *ValEase_4R_S1*, *ValEase_5_S1*). Component 3 had high loadings from 4 ease-of-use affordance items (*AffEase1_S1*, *AffEase_2_S1*, *AffEase_3_S1*, *AffEase_6R_S1*). Component 4 comprised 2 ease-of-use affordance relevance items (*RelEase_3_S1*, *RelEase_4_S1*).

The four-factor structure, with distinct affordance, relevance, valence, and usefulness/use intention components, suggested participants understood items within each scale similarly while distinguishing between scales.

Table 6. 3 Pattern Matrix with 4 Components on Experiential Variables for Session 1 in Study 2

Items	Component			
	1	2	3	4
AffEase_1_S1			.853	
AffEase_2_S1			.797	
AffEase_3_S1			.851	
AffEase_6R_S1			.698	
RelEase_3_S1				.790
RelEase_4_S1				.818
ValEase_1_S1		.708		
ValEase_2_S1		.714		
ValEase_3R_S1		.870		
ValEase_4R_S1		.775		
ValEase_5_S1		.833		
UseIntR_1_S1	.855			
UseIntR_2_S1	.820			
UseIntR_3_S1	.765			
UseIntR_5_S1	.855			
UseP_1_S1	.933			
UseP_2R_S1	.686			
UseP_3_S1	.847			
UseP_4R_S1	.908			
UseP_5_S1	.922			

6.3.1.3 Outliers Exploration and Mean Calculation

Next, possible acquaintance bias on three **Medium** for each scale was checked. In the *Text* sessions, participant#146 had contradictory answers for items *AffEaseT_4R*; participant#162 for *UsePT_2R* and *UsePT_4R*; and participant#157 for *ValEaseT_3R* and

ValEaseT_4R. In the *Audio* sessions, participant#162 and participant#131 had contradictory answers for items *AffEaseA_4R*, participant#66 in *UsePA_4R* and *ValEaseA_4R* items, and participant#140 in *RelEaseA_3* item. Regarding the *Video* sessions, participants#148, #172 and #156 had contradictory answers for counter-indicators items like *AffEaseA_4R* and *AffEaseA_6R*; participant#157 had higher rating on *RelEaseV_3* and *RelEaseV_4*; and participant#131 had contradictory rating on *ValEaseV_3R* and *ValEaseV_4R* items to other items.

Specifically, the participants only had contradictory rating in only one or a maximum of two scales (participant#66 in *Audio* session) in their questionnaire per session. Most of the contradictory items were counter indicators. This suggested no overall acquaintance bias over the questionnaire existed, and the contradictory items that occurred might be due to carefulness. Also, some contradictory items, e.g., *AffEaseA_4R*, were not included into the analysis after the reliability and validity analysis. Therefore, no case was disregarded for further analysis due to acquaintance bias.

Following this, the mean values for all items in the five remaining scales in three **Mediums** were calculated. The 15 mean values were distinguished from the single-item values, using the prefix *M_*, for example, *M_AffEase#_1*. The # in the example represents **Medium** with *T* representing "*Text*", *A* representing "*Audio*" and *V* representing "*Video*". After Boxplot exploration for each scale in each Medium, eight below-average outliers occurred for one or more scales, as depicted in Table 6.4. Specially, participant#140, #150, #157 and #158 had extreme mean values for more than one scale. To avoid messing up the experience across **Medium**, these four cases were disregarded for group comparison of the experience and path analysis. Therefore, 36 participants ($N = 36$) included in group comparison and path analysis.

Table 6. 4 Outlier Distribution across Medium in Study 2

Medium	Text	Audio	Video
M_AffEase	(131, 161)	/	/
M_RelEase	(140)	(150, 157)	(116, 140)
M_ValEase	/	(157)	(140)
M_UseIntR	/	(140, 150, 157, 158)	(140, 150, 157)
M_UseP	(140, 157)	(122, 140, 157, 158)	(140, 157*)

Note. Outliers in the lower end of the box are marked with a bracket.

6.3.2 Qualitative Data

6.3.2.1 Coding Results

Thematic analysis was used to analyze the interview data. A codebook was developed based on the interview transcripts, resulting in 75 codes organized into a hierarchy with the following top-level categories:

1. Background, encompassing codes related to participant identity, occupation/major, emotional distress, diagnostic method, emotional distress triggers, emotion coping mechanisms, and robot acceptance.
2. Exercise evaluation, including codes for reasons to attend, understanding of the exercise, effect expectations, coercive offenses, and individual needs.
3. Medium differences, covering codes related to concentration, companionship, interactivity, conviction, oppression, necessity, and naturalness.
4. Online therapy selection criteria, with codes for standardized exercises, counseling context, trust, and needs.
5. Trust based codes related to function, relation, emotion, privacy, and effect.

For coding, a participant utterance was defined as one turn taken in conversation, for a total of 2163 utterances. Each utterance was coded holistically and could be assigned more than one code.

Ideally, the utterances should have been coded by three coders and measured for inter-rater reliability using Cohen's kappa (McHugh, 2012). However, due to time constraints, the researcher coded the results on her own at different time points to compare and adjust the codes.

The results section (6.4) interprets the emergent themes relevant to answering the research questions rather than all the coding results. Throughout the results, numbers in parentheses indicate how many participants explicitly supported a given point. If a participant is not counted, it does not necessarily mean they disagreed, but rather that they did not explicitly mention that point.

6.4 Results

6.4.1 Quantitative Analysis Results for Research Questions

6.4.1.1 Experiential Variables Comparison among Medium

As most data were non-normally distributed, nonparametric Friedman tests compared the three paired samples (*Text*, *Audio*, *Video*) on the experiential variables. The Friedman test, a nonparametric alternative to repeated Analysis of variance (ANOVA) measures, is appropriate when the same participants are measured under three or more conditions. The null hypothesis (H0) is that the distribution of scores is the same across conditions. Rejecting H0 at $p < .05$ indicates differences between at least two conditions.

Results (Table 6.6) showed a significant difference in *M_UseIntR* (use intention to robot) across **Medium**, $\chi^2(2) = 8.75$, $p = .013$. Bonferroni-corrected Wilcoxon signed-rank post-hoc tests (Table 6.7) revealed higher use intention in the *Audio* (M = 4.569) versus *Text* condition ($M = 4.097$), $Z = -2.65$, $p = .024$. No other significant differences emerged between **Medium** on the remaining experiential variables (*M_AffEase*, *M_RelEase*, *M_ValEase*,

M_UseP) manipulation checks (*Check_EngImg*, *Check_EaseImg*, *Check_EngExcer*, *Check_SatR*, *Check_ExpR*) at the adjusted significance level.

In summary, the Friedman test indicated a difference between **Medium** specifically on use intention, with participants reporting greater intention to use the audio robot compared to the text chatbot.

Table 6. 5 Related-Sample Friedman's Two-Way Analysis of Variance by Ranks in Study 2

Measures	Medium (Mean Rank)			Friedman's test statistic	Asymptotic Sig.	N
	Text	Audio	Video			
Experiential variables						
M_AffEase	2.03	2.14	1.83	2.26	.324	36
M_RelEase	1.85	2.15	2.00	2.28	.319	36
M_ValEase	2.04	1.97	1.99	0.11	.948	36
M_UseIntR	1.72	2.35	1.93	8.75	.013	36
M_UseP	1.83	1.97	2.19	2.71	.258	36
Manipulation checks						
Check_EngImg	1.75	2.08	2.17	4.94	.085	36
Check_EaseImg	1.81	2.01	2.18	3.33	.189	36
Check_EngExcer	1.83	2.15	2.01	2.63	.268	36
Check_SatR	1.83	2.04	2.13	2.32	.314	36
Check_ExpR	2.07	1.95	1.98	.317	.853	30

Table 6. 6 Pairwise Comparisons with Wilcoxon Signed-Rank Tests in Study 2

Sample 1 (Mean, SD)		Sample 2 (Mean, SD)	Sts. Test Statistic (Z)	Adj Sig. (adjusted by Bonferroni correction)	N
M_UseIntR_T (4.097, 1.125)	-	*M_UseIntR_A* (4.569, .677)	-2.652	0.024	36
M_UseIntR_T (4.097,1.125)	-	*M_UseIntR_V* (4.403, .0728)	-.884	1.000	36
M_UseIntR_V (4.403, .0728)	-	*M_UseIntR_A* (4.403, .0728)	1.768	0.231	36

6.4.1.2 Experiential Variables Path Analyses Moderated by Medium

A moderated mediation model (Figure 6.1) was tested using PROCESS (Hayes, 2018; model 58; 5,000 bootstraps; $N = 108$). The model had *M_AffEase* as the independent variable, *M_UseIntR* as the dependent variable, *M_RelEase* and *M_ValEase* as mediators, and **Medium** as the moderator of the indirect effects.

It test the following hypotheses: **H2.1** *M_AffEase* has a significant direct influence on *M_UseIntR*;**H2.2** *M_AffEase* has a significant direct influence on *M_RelEase*;**H2.3** *M_RelEase* has a significant direct influence on *M_UseIntR*; **H2.4 Medium** moderated the effect of *M_AffEase* on *M_RelEase*; **H2.5 Medium** moderated the effect of *M_RelEase* on *M_UseIntR*; **H2.6** *M_RelEase* mediates the effect of *M_AffEase* on *M_UseIntR*; **H2.7 Medium** moderated the mediated effect of M_RelEase on the path of *M_AffEase* to *M_UseIntR*; **H2.8** *M_AffEase* has a significant direct influence on *M_ValEase*; **H2.9** *M_ValEase* has a significant direct influence on *M_UseIntR*; **H2.10 Medium** moderated the effect of *M_AffEase* on *M_ValEase*; **H2.11 Medium** moderated the effect of *M_ValEase* on *M_UseIntR*; **H2.12** *M_ValEase* mediates the effect of *M_AffEase* on *M_UseIntR*; **H2.13 Medium** moderated the mediated effect of *M_ValEase* on the path of *M_AffEase* to *M_UseIntR*.

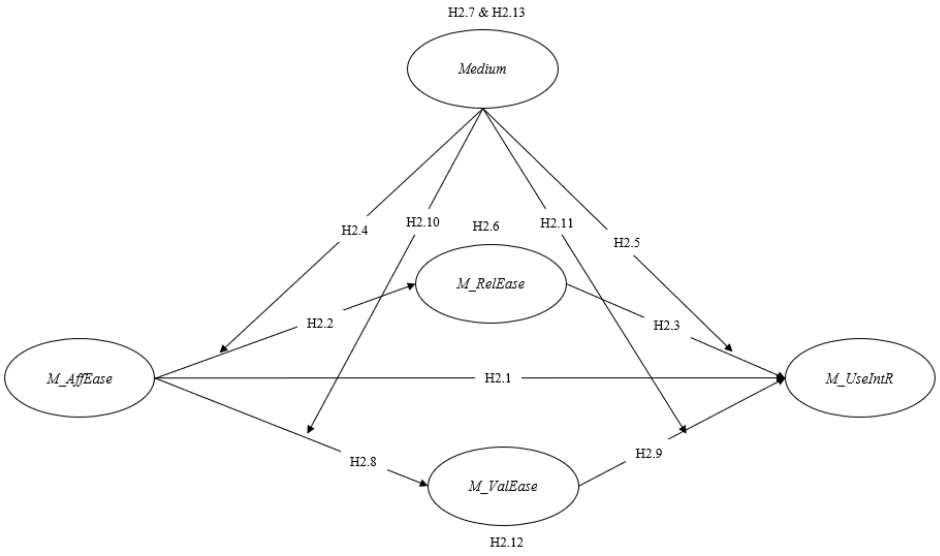

Figure 6.1 The Model Framework and Hypotheses Model Testing.

Results (Figure 6.2) showed a significant direct effect of *M_AffEase* on *M_RelEase* (*B* = .5462, *SE* = .2644, *p* = .0413) but not on *M_ValEase* (*B* = .3577, *SE* = .2676, *p* = .1843). The direct effect of *M_AffEase* (*B* = -.0367, *SE* = .0899, *p* = .6840), *M_RelEase* (*B* = .6585,

SE = .2662, *p* = .0150) and *M_ValEase* (*B* = .6138, *SE* = .2832, *p* = .0326) on *M_UseIntR* were nonsignificant. **Medium** did not moderate any direct effects.

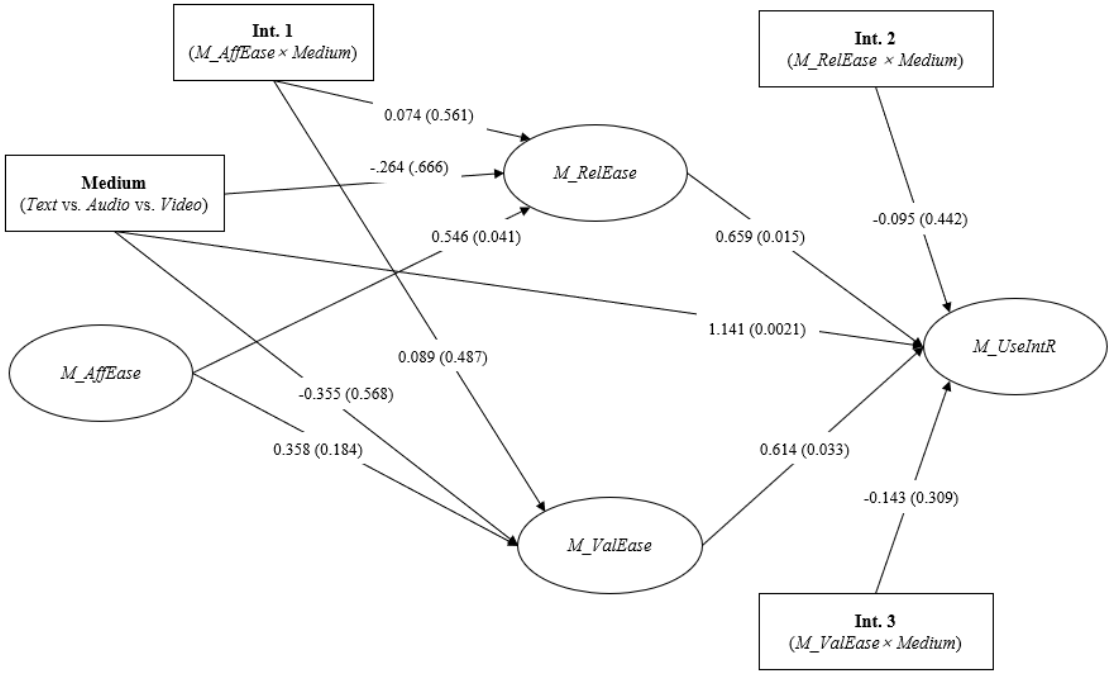

Figure 6.2 The Moderated Effect of Medium on the Relationship Between M_AffEase and M_UseIntP Through M_RelEase and M_ValEase with Estimated Coefficients (bootstrap) for N = 36: Significance value was in bracket.

The indirect effect of *M_AffEase* on *M_UseIntR* through *M_RelEase* was significant for all **Medium** conditions: *Text* (*Effect* = .3489, *BootSE* = .1333, 95% CI [.0916, .6163]), *Audio* (*Effect* = .3241, *BootSE* = .0893, 95% CI [.1349, .4843]), and *Video* (*Effect* = .2853, *BootSE* = .1417, 95% CI [.0140, .5617]). The indirect effect through *M_ValEase* was significant for *Text* (*Effect* = .2106, *BootSE* = .1016, 95% CI [.0213, .4181]) and *Audio* (*Effect* = .1764, *BootSE* = .0674, 95% CI [.0580, .3228]), but not *Video* (*Effect* = .1168, *BootSE* = .1255, 95% CI [-.1167, .3799]). *M_RelEase* fully mediated and *M_ValEase* conditionally mediated the effect of *M_AffEase* on *M_UseIntR*.

The results suggest that ease-of-use affordance indirectly influenced robot use intention (rejecting H2.1) by increasing relevance (supporting H2.2, H2.3, H2.6), regardless

of **Medium** (rejecting H2.4, H2.5, H2.7). This underscores the role of affordance-goal congruence (task-fit affordance) in shaping acceptance, consistent with the I-PEFiC framework.

Valence of ease-of-use interaction mediated the effect for *Audio* and *Text*, fully supporting H2.13 but partially supporting H2.8, H2.9 and H2.12. However, **Medium** did not significantly moderate the indirect effects through valence (rejecting H2.10 and H2.11). This suggests robot features may amplify/attenuate valence beyond affordance-goal comparisons alone, highlighting the need to explore contextual factors shaping valence in the I-PEFiC model. The interview data can provide qualitative insights into variability in experiential processing and inform human-robot interaction design.

6.4.1.3 Correlation of M_UseIntR and M_UseP

According to I-PEFiC, robot use intention (*M_UseIntR*) influences overall system usefulness evaluations (*M_UseP*), but as a recurrence processing model, usefulness appraisals also continually shape users' intentions. Rather than examining causality, I focused on whether *M_UseIntR-M_UseP* correlation differed across **Medium** using Cohen's Q test.

Pearson correlations were strong for Text ($r1 = .724$), *Audio* ($r2 = .568$), and *Video* ($r3 = .645$), all $ps < .001$. Fisher's r-to-z transformation yielded $z1 = 0.92$, $z2 = 0.64$, and $z3 = 0.77$. The overall Q statistic (1.931) did not exceed the critical value of 5.991 with $df = 2$ at $\alpha = .05$. Pairwise comparisons (*Text* versus *Audio* $Q1 = 1.285$; *Text* versus *Video* $Q2 = .369$; *Audio* versus *Video* $Q3 = .277$) were also nonsignificant at $df = 1$ and $\alpha = .05$, providing no evidence of correlation differences across **Medium**. Collapsing across **Medium**, the overall *M_UseIntR-M_UseP* correlation was $r = .657$ ($p < .001$), indicating a strong positive relationship regardless of robot modalities.

6.4.1.4 Summary

The quantitative results provide initial evidence for **RQ2.1**. The lack of significant differences on most experiential variables suggests comparable perceptions between the *Text*, *Audio*, and *Video* robots among depressed individuals. However, the higher use intention for the audio bot than the text chatbot indicates some divergence.

The path analysis highlights the consistent mediating role of relevance evaluations across robots, while the conditional indirect effect (only in the *Audio* and *Text* groups) via valence implies boundary conditions around robot features amplifying valence beyond affordance-goal comparisons alone.

The correlation analysis showed a significant moderate relationship between robot use intention and **eiIBM_RobotV1** usefulness evaluation, suggesting ease-of-use affordance does not fully determine overall program usefulness perceptions.

The next section's qualitative insights can provide a richer understanding of how specific attributes differentially shape depressed individuals' experiences, particularly regarding valence formation and acceptance determinants beyond task-fit affordance, helping to elaborate the quantitative findings.

6.4.2 Qualitative Analysis Results for Research Questions

The qualitative analysis explored participants' robot perspectives by answering four interrelated questions:

1. What significant evaluations of the different robots emerged from the interviews? (Qualitative question 1)
2. Which factors shaped participants' preferences and avoidances for specific robots? (Qualitative question 2)

3. How do emotional distress triggers and expectations relate to robot selections? (Qualitative question 3).
4. What are participant-inspired insights for improving perceptions of effectiveness and engagement with the **eiIBM_RobotV1** program? (Qualitative question 4)

Sequentially exploring these questions provides a comprehensive multilayered understanding, highlighting key robot evaluations, their linkages to preferences, the role of individual differences, and participant-inspired recommendations to enhance experiences. Together, the analyses offer key considerations and opportunities around social robots for emotional support.

6.4.2.1 Qualitative question 1: What significant evaluations of the different robots emerged from the interviews?

The thematic analysis of the evaluation of the robots identified seven key subthemes characterizing how the different agents impacted participants' experience (Theme 1: evaluation of the medium; Table 6.7).

Concentration theme (subtheme 1.1) highlighted how the robot's features helped or hindered focus on the exercise. Most participants (19) noted how the audio modality enabled easier engagement compared to potential visual distractions with the video modality or reading text. However, the telepresence robot's behavior or appearance distracted some participants (8) from exercise engagement, with two attributing this to its machine-like physical appearance and one to the small screen size. For the chatbot, opinions were divided, with 9 participants reporting that the text helped them focus, while 8 found the heavy reading boring and difficult to concentrate on. Four of these participants mentioned a preference for reading text.

Companionship theme (subtheme 1.2) showed how the robot provided a sense of social support during the exercise. The audio bot elicited feelings of chatting with a friend for 18 participants, and the telepresence robot's physical embodiment contributed to perceived companionship for 13 participants. While one participant found the chatbot cold and automated, another likened it to chatting with friends.

Oppression theme (subtheme 1.3) revealed how the agent could potentially overwhelm participants. Some felt pushed to complete the exercise with the audio robot (1 participant) and chatbot (2 participants), while 3 participants from the *Text* group found it tiring. The telepresence robot provoked anxiety in 2 participants who reported under pressure of the robots.

Interactivity theme (subtheme 1.4) emphasized the interaction level during the exercise, with 6 participants mentioning this evaluation for telepresence robots. No one mentioned this aspect for the chatbot or audio bot.

Conviction theme (subtheme 1.5) showed how the robot's features affected the perceived believability in its interpretation. The stiff expressions or apperance of the telepresence robot (7 participants) and audio bot (1 participant) reminded them it was not human, reducing the convincingness of their encouragement or interpretation. However, 2 participants found the audio bot's interpretation convincing. The chatbot did not elicit any conviction-related evaluations.

Unnecessity theme (subtheme 1.6) demonstrated perceptions of indispensable versus unneeded robot features. Nine participants saw the telepresence robot's movements and appearance unrelated to the task as superfluous. Two participants, who also expressed a preference for reading text, reported that sound from the audio bot was unnecessary for the task.

Finally, unnaturalness theme (subtheme 1.7) highlighted how robotic tone or synchronization impacted perceptions of humanness. Stiff, automated speech and movements were seen as hindering humanness, with most mentions relating to the telepresence robot (7), followed by the audio bot (2). This could be due to the decreasing anthropomorphic characteristics of these robots. Although participants found the chatbot (3)'s text message unnatural, this was attributed to the individual's habits of using Cantonese characters. When accompanied by audio, the confusion might be reduced as people might not fully grasp the message from the text alone.

Taken together, the thematic analysis revealed nuanced evaluations for each robot modality. For the telepresence robot, key evaluations centered around concentration, companionship, unnecessity of physical features, unnaturalness, interactivity, conviction, and oppression. Participants focused on how the robot's embodiment and motions impacted various aspects of their experience. In contrast, evaluations of the audio bot primarily involved concentration and companionship, with participants concentrating on how the voice impacted their experience without physical embodiment. For the text chatbot, core evaluations revolved around concentration and oppression through the demands of reading, with less commentary on companionship or anthropomorphism in the absence of physical or vocal cues.

Table 6. 7 Subthemes of Significant Evaluation of Robots Identified in Transcripts

Evaluation	Key Reasons and subthemes (no. of participants)	Examples
	Chatbot	
Concentration	(+) Single stimulus improves focus (3)	*"For me, the text-only version... If you have to pay attention to audio and visuals, it's easy to get distracted. My concentration is not that good."*
	(+) More imaginative space (1)	*"...because it (text) has the strongest active imagination. Different people may have different preferences. For me, imagining from text is a bit better."*
	(+) Familiar (1)	*"Texting is familiar for me, like chatting with friends."*
	(+) Less likely to miss words than verbal information (1)	*"With text, I can complete the task faster... When it (audio) describes the situation, I might miss a sentence and have to think back. But with text, I can look back."*
	(+) General concertation (3)	*"For focus, the third one, the pure text version, was best."*

	(-) Difficulty in focus on reading (4)	*"Because for the first two (text and audio), I would get a bit distracted."*
	(-) Difficulty in imaging the text (2)	*"With pure text, it may be harder to immerse (into the described scenarios)."*
	(-) confusion due to oral expression in text (2)	*"The text you (the robot) type is colloquial… a bit strange when I read some of the words"*
Companionship	(+) like chatting with friends (1)	*"Texting is familiar for me, like chatting with friends."*
	(-) cold text, like automative response (2)	*"Just text feels a bit cold... like an automatic reply."*
Oppression	(-) felt pressured by the reminders to fill out the ending (1)	*"Sometimes, when I haven't thought of the word yet, it quickly urges me to answer."*
	(-) tiring (2)	*"Reading text can sometimes feel very tiring or sleepy. I want to skip some words. I have to concentrate to read all the text."*
	(-) pressure, like doing an examination (1)	*"If it's pure text, it feels like doing an exam exercise."*
Unnaturalness	(-) expressions different from my way of speaking (3)	*"Some colloquial wordings in text may not be what we usually type, which feels a bit strange."*
	Audio bot	
Concentration	(+) Easey to follow (5)	*"(with audio), I actually feel it allows me to enter the situation faster and concentrate more on doing it...*
	(+) Emotional sound (1)	*"If there is sound, it would be more focused…Although it's read by AI, it still feels like there are some emotions when you listen to it"*
	(+) Save mental effort with audio (1)	*"It's like having someone accompany me to do it… if I read the text myself, it might be a bit more mentally exhausting."*
	(+) General felt concentrated (12)	*"It (audio) can be more immersive than text"*
Companionship	(+) Like chatting with a friend (1)	*"It feels more like chatting than text."*
	(+) Have someone leading me (1)	*"it will lead you to complete the whole task... there is a response to your answer"*
	(+) Synchronization with the task (1)	*"(With audio), at least I know it is on the same progress with me."*
	(+) Encouraged by the sound (1)	*"It (audio) will really verbally tell you if you got the answer right. But if you read text, there isn't much of this encouraging feeling."*
	(+) General felt companioned (14)	*"If it's about the feeling of companionship, I think audio is better."*
Oppression	(+) Felt like it was a listening comprehension test (1)	*"It feels a bit like a listening comprehension exercise."*
Conviction	(+) General felt it convincing because of reasonable (2)	*"Listening to a human voice (from audio robot) makes you feel a bit more humanized."*
	(-) Stiff expression makes me notice it is robot (1)	*"The audio highlighted some characteristics, letting me know it is a robot. Some wrong pronunciation emphasis makes me realize it is a robot."*
Unnecessity	(-) low efficiency, have to wait for reading or performing (2)	*"It speaks slower than I read. I have to wait …"*
Unnaturalness	(-) remined me of its machine characteristics (2)	*"The voice is mechanical… though describing the scenario, it does not make difference for me"*
	Telepresence robot	
Concentration	(+) Minimal distraction caused by its slight movement (2)	*"Its movements are suitable and simple. You can focus on it without being distracted… listening to it to finish the task does require me much attention energy."*
	(+) Less faith in effect so it did not distract me (1)	*"It's (the audio and visual characteristic) a minor and least important thing for me. I didn't find the exercise very useful, so it didn't negatively impact me."*
	(+) General felt concentrated (1)	*"Robot helps me focus more…"*
	(-) Difficulty engaging due to toy/machine-like appearance (2)	*"It looks like a toy robot, so I might not take it very seriously…"*
	(-) Distracted by its appearance and behaviors (4)	*"It is quite interesting to watch its movements, but it probably doesn't help me focus on the imagined scenario."*
	(-) Distracted because of the screen overlap the text (1)	*"I have to type and adjust the screen to see the full text, which is not very convenient."*
	(-) Pressure to watch robot thus could not focused (1)	*"The robot was almost frightening… I don't really like it but fear it."*
Companionship	(+) Notice the embodiment (1)	*"I can see something on the screen accompanying me."*
	(+) General sense of companionship (9)	*"I really feel like there's something accompanying me"*
	(-) looking cool (1)	*"It's very aloof, talking to a robot like that... I know it's a robot, but it makes me a bit unhappy."*

	(-) The feedback is cold like automated response (1)	*"The robot is set with a formula, it will respond with the same thing when it sees the answer."*
	(-) too small to see clearly (1)	*"The robot is small, so it... I mean, it's not very clear to see."*
Oppression	(-) scared of watching robots (1)	*"It was quite frightening… and it's not very nice looking"*
	(-) felt pressure when watching the robot (1)	*"Mainly because looking at the robot, I actually feel quite unhappy."*
Interactive	(-) not much interaction as expected (1)	*"Seeing the robot, I felt it didn't particularly have rich expressions or movements, not particularly engaging."*
	(+) compared to other robots, it has more interaction with me (5)	*"I feel like I'm actually interacting with something. And if it was just playing the sound and showing the text dialogues, I would feel the interaction is not very strong."*
Conviction	(-) Appearance reminded me of it being a robot (4) and preferred avatar (2 out of 4)	*"Because it's too cartoonish, it then becomes less convincing."*
	(-) Appearance remined me of it being robot and it is silly to say to a robot (1)	*"…I feel it's silly to talk to a robot."*
	(-) The stereotype of robot prevented me from believing in it (1)	*"You might have a preset impression that it (the robot) won't understand you."*
	(-) The stiff expression makes me notice it is a robot (1)	*"The audio highlighted some characteristics, letting me know it is a robot. Some wrong pronunciation emphasis makes me realize it is a robot."*
Unnecessity	(-) Movement do not correlated to content (1)	*"I feel that its movements aren't very relevant to what its saying."*
	(-) Appearance do not correlated to content (2)	*"So you might see its appearance, but its appearance doesn't really help with the exercise."*
	(-) Apart from the other component – text and image (1)	*"You see the robot's appearance there, but it doesn't show the scenario or image in the video call…"*
	(-) No contribution to exercise (4)	*"The purpose of a WhatsApp call is that I can speak directly. But I have to type at the same time… if so, I'd rather not look at the screen."*
	(-) Low efficient, I have to wait for its performing (1)	*"I feel like I have to wait for the robot to finish reading the words, and maybe the robot…"*
Unnaturalness	(-) response so slowly, reducing interaction feeling (1)	*"I wait for it to finish speaking, and then look at the text, it actually feels very slow."*
	(-) do not attractive in movement or expression (2)	*"I feel it doesn't have particularly rich expressions or movements, not particularly engaging"*
	(-) General sense of stiff tone and movement (4)	*"The robot's voice is a bit stiff, because after all, it's not human, so its emotional aspect or the tone of speech might not be so authentic."*

6.4.2.2 Qualitative question 2: Which factors shaped participants' preferences and avoidances for specific robots?

While 7 participants reported equal preference for the different robots, the majority expressed a slight to strong preference for specific robots over others. The audio bot was the most preferred (18 participants), followed by the telepresence robot (9 participants, one of which showed a similar preference for the audio bot), and the chatbot (5 participants, one of which also reported equal preference for the audio bot). The extent to which the evaluations contributed to robot preferences differed. Table 6.8 summarizes the connections between

positive evaluations (themes) and preferred robots, while Table 6.9 indicates how negative evaluations contributed to avoiding certain robots.

6.4.2.2.1 Positive Evaluation Contribution

The audio bot's concentration benefit was praised by 18 participants for enabling better focus, with 14 of them selecting it as their most preferred robot. The companionship benefit was reported by 17 participants, who felt like someone was leading them and encouraging them, with 12 of those participants selecting it as their preferred robots. Conviction also contributed to 2 people reporting the audio bot as their preferred choice, although these participants also found the audio bot helped with concentration and provided a sense of companionship.

The telepresence robot's companionship was valued by 10 participants for providing embodied presence during the exercise. Of those 10, 5 chose it as their top preference, suggesting that its physicality was pivotal for many favoring this modality. The telepresence robot's greater interactivity was positively noted by all 5 participants who ranked it as their first preferences, indicating that physical interactivity was crucial for those preferring this robot.

The text-based chatbot's concentration benefit through reading was discussed by 9 participants, but only 3 of them chose it as their most preferred robot, suggesting that concentration alone, without additional modality influences, does not strongly sway preference for the chatbot.

Table 6. 8 Evaluation Determinants of Robots Contributing to Preference on the Robot

Preference on the Mediums	**Evaluation Aspects (Positive)**			
	Concentration	Companionship	Interactivity	Conviction
Chatbot (5)	3/9	0/1	/	/
Audio bot (19)	14/18	12/17	/	2/2
Telepresence robot (9)	3/4	5/10	5/5	/
No preference (7)	1/3	/	/	/

Note. The extent of the contribution is highlighted by color, the deeper the grey, the more important the evaluation for preference robot selection.

6.4.2.2.2 Negative Evaluation Contribution

The chatbot's reading oppression was described by 4 participants as causing overwhelm, with 3 of them choosing the audio bot and 1 choosing the telepresence robot instead, indicating that reading demands deterred its selection. Another 7 participants did not choose the chatbot due to difficulties in concentrating on reading the text, leading 4 of them to choose audio bot and 2 to choose the telepresence robot. Although one participant (#132) reported a lesser sense of companionship from the chatbot, they still chose it as their preferred robot.

The telepresence robot's physical distractions were reported by 9 participants as disruptive to focus, with 6 of them ranking the audio bot higher and 1 ranking the chatbot higher, demonstrating that its disruptiveness discouraged its choice. The unnecessity of certain features in the telepresence robot was also criticized by 9 participants, with 6 of them ranking the audio bot or chatbot higher. Another criticism was the lack of conviction within the robot's physical appearance, with 6 participants reporting its toy-like appearance made the exercise less serious and convincing, leading them to prefer the audio bot, which did not show its appearance. Six participants also criticized the telepresence robot's unnaturalness, although 2 of these critics still preferred it. Three other participants avoided the telepresence robot, stating that it was pressuring or unpleasant to look at the robots (oppression subtheme).

The audio bot received the least negative evaluation, with only 5 participants finding it to have unnecessity function (2 participants), unnaturalness (2 participants), oppression (1 participant), less concentration (1 participant), and less convincing (1 participant). However, the participant (#116) who found the audio bot 's tone stiff and its less convincing word still

chose the audio bot as their preference because it helps them focus and provided companionship.

Table 6. 9 Evaluation Determinants of Robots Contributing to Avoidance from the Robot

Preference on the Mediums	Evaluation Aspects (Negative)					
	Unnecessity	Unnaturalness	Oppression	Concentration	Companionship	Conviction
Chatbot (5)	/	Audio bot: 1/1	Audio bot: 3/4. Telepresence robot: 1/4	Audio bot: 4/7 Telepresence robot: 2/7	Chatbot: 1/1	/
Audio bot (19)	Chatbot: 1/2. No preference: 1/2.	Audio bot: 1/2. Chatbot: 1/2.	Telepresence robot: 1/1	Chatbot: 1/1		Audio bot: 1/1
Telepresence robot (9)	Audio bot: 5/9. No preference: 2/9 Telepresence robot: 1/9. Chatbot: 1/ 9	Audio bot: 3/6. Telepresence robot: 2/6. Chatbot: 1/6	Audio bot: 1/3. No preference: 1/3	Audio bot: 6/9. Chatbot: 1/9 Telepresence robot: 1/9. No preference: 1/9	No preference: 1/1	Audio bot: 5/6. Chatbot: 1/6
No preference (7)	/	/	Telepresence robot: 1/1			

Note. The extent of the avoidance contribution is highlighted by color, the deeper the orange, the more important the evaluation for preference robot deselection.

6.4.2.2.3 Summary

The key factors determining participants' preferences for, or avoidance of specific robots are evident. The audio bot was most preferred largely due to its strengths in sustaining concentration and conveying a sense of companionship, with its minimal physical form limiting potential distracting features. In contrast, the telepresence robot's physicality was pivotal for some but a detriment for others, as its embodiment enabled interactivity and companionship yet also posed visual distractions. Those troubled by distracting features or questioning the robot's necessity often ranked the audio bot or chatbot higher. The chatbot appealed to a smaller subset valuing the concentration benefit of reading over other relational or embodied aspects. However, its singular text interface frequently felt overwhelming or inadequately engaging for many.

These findings highlight the importance of aligning robot capabilities with individual priorities and needs. However, a question emerges regarding whether robot preferences stem from innate personal requirements or different evaluative focuses for specific robots. For

example, only a few participants (2 out of 8) who valued the chatbot's concentration benefit ultimately preferred it, introducing the quandary of whether chatbot selection is driven by concentration being a foremost priority for those individuals, or if concentration is a secondary need only served by the chatbot when other priorities like companionship cannot be adequately fulfilled by alternate robots. In other words, are robot selections shaped by the positive valence of met expectations for one robot, or the negative valence of unmet expectations for other robots? Further exploration of the relationships between individual expectation needs, emotional distress triggers, and robot preferences would elucidate this issue. Clarifying the motivators underlying users' robotic preferences and selections can inform the design of human-robot interactions that effectively align technological capabilities with the nuanced priorities of target user groups.

6.4.2.3 Qualitative question 3: How do emotional distress triggers and expectations relate to robot selections?

6.4.2.3.1 Emotional Distress Trigger and Preference of Robot

Depression is a complex mental health condition that can be triggered by a combination of tangible and intangible factors. Tangible triggers refer to external events or circumstances that contribute to the development or exacerbation of depression, such as stressful life events or traumatic experiences. Intangible triggers are more internal or psychological factors (Ponte, 2022), such as negative thought patterns or unreasonable negative mood. Table 6.10 lists the coding of participants' emotional distress triggers with examples.

Among the 19 participants preferring the audio bot, 12 had tangible triggers, while 7 had intangible triggers. Of the 9 participants preferring the telepresence robot, 7 had tangible triggers and 2 had intangible triggers. Among the 5 participants preferring the chatbot, 3 had

tangible triggers and 2 had intangible triggers. It appeared individuals with observable distress sources tended to prefer the audio bot (12 out of 18), while those with internal triggers were about evenly split between preferring the audio bot (7) and telepresence robot (7).

Table 6. 10 Emotional Distress Trigger Type across Robot Preference

Preference on the Mediums	Emotional Distress Triggers	
	Tangible triggers	Intangible triggers
Chatbot (5)	3	2
Audio bot (19)	12	7
Telepresence robot (9)	2	7
No preference (7)	4	3

Surprisingly, 11 out of 16 participants with tangible triggers (excluding four who did not have a robot preference) chose their preferred robot due to concentration benefits. In contrast, 10 out of 16 participants with intangible triggers (excluding three without preferences) selected their preferred robot based on companionship attributes. This indicates that robot designs for individuals with tangible versus intangible triggers should emphasize concentration and companionship features respectively.

6.4.2.3.2 Individual Expectations of Exercise

Next, participants' expectations for **eiIBM_RobotV1** were examined. Most participants did not report any expectations after participating or felt unsure how robots could better assist them in the exercise, whether in terms of exercise content or robot capabilities. About half explicitly expressed their expectations or needs, categorized into three aspects: response to the possible contradict resolution to the scenario (narrow resolution response); functional expectation around training effect (functional need); and emotional support needs including emotional relief and empathy from the robot (emotional need).

Regarding narrow resolution response, 8 participants expressed feeling offended or uncomfortable, while 12 did not. Sources of offense included being pressured to correct errors (4 participants), difficulty believing unrealistic positive responses (2 participants), and desiring respect for multiple perspectives (2 participants). Those not offended viewed it as merely exercise (10 participants) or beneficial practice (3 participants). Comparing trigger types, the 4 unoffended participants with intangible triggers all chose the telepresence robot, while tangible triggered participants chose non-telepresence robots, saying they did not want to dispute with the machine.

For functional needs, 9 participants described cognitive expectations around the **eiIBM** practice. such as challenging thinking (2), providing attributions explaining scenario resolutions (2), teaching acceptance rather than pushing beliefs (1), and desiring slower pacing, more interaction, and time to adjust thoughts (4).

Regarding emotional support needs, 13 participants expressed expectations, including help shifting attention and reducing life distress (9), changing their situations (2), and more humanized, empathetic responses from the robot (2).

Further analysis revealed connections between participants' expectations, emotional triggers, and robot preferences (Table 6.11). Those with intangible triggers who expected greater emotional support and interactivity overwhelmingly preferred the telepresence robot (4 out of 4). They also seldom reported being offended by the narrow resolution answers. In contrast, the most tangible triggered participants desiring emotional support chose the audio bot (3 out of 4).

Table 6. 11 Individual Expectation of Exercise across Robot Preference

Preference on the Mediums (No. of participants)	Individual Expectations of Exercise			
	Narrow resolution response	Function need	Merger of functional need and emotional needs	Emotional Needs

	Offended	No Offended	Mindset & Perspective	Personal Experience & Beliefs	Program Interaction & Flexibility	Emotional & Psychological Support	Humanization & Empathy
Chatbot (5)		1/0	0/1	/	/	1/1	/
Audio bot (19)	3/1	5/2	1/1	1/0	1/0	3/0	0/1
Telepresence robot (9)	1/1	0/4	/	/	0/3	0/4	/
No preference (7)	0/1	2/0	0/1	/	//	1/1	/

Note: the number on the left in each cell represents the number of participants with tangible triggers, and the number on the right represents the number of participants with intangible triggers.

6.4.2.3.3 Summary

The findings reveal connections between participants' emotional distress triggers and their robot preferences. Those with tangible triggers tended to select audio bots, likely because the concentration support aligned with their need to address defined concerns. In contrast, participants with intangible triggers often preferred telepresence robots that provided companionship to escape negative moods lacking clear sources. Alignment with psychological needs explains these patterns - tangible triggered participants prioritized concentration to resolve triggers directly, while intangible triggered participants valued companionship for temporary relief without resolving trigger causes.

Additionally, robot preferences aligned with individual expectations shaped by emotional triggers. Participants with intangible triggers overwhelmingly chose telepresence robots if desiring greater emotional support and interactivity, as its physical embodiment seemingly satisfied their expectations. Conversely, tangible triggered participants seeking emotional help often opted for the audio bot's concentration. The chatbot appealed most to those valuing cognitive over emotional engagement. These findings highlight the importance of tailoring robot capabilities and attributes to match the distinct expectations and needs arising from unique depression triggers.

6.4.2.4 Research question 4: What are participant-inspired insights for improving perceptions of effectiveness and engagement with the eiIBM_RobotV1 program?

Perceived usefulness is critical for continued technology acceptance and use, so it is important to understand how to improve the **eiIBM_RobotV1** implementation to enhance trust in its effectiveness. The subthemes that emerged from the evaluation of the program's effectiveness and the reasons described are listed in Table 6.12.

Thirty participants reported their opinion on the effectiveness of **eiIBM_RobotV1**. Fourteen had a positive valence, believing the exercise trains their subconscious negative patterns (6; e.g., "the robot's ending different from mine but logically valid, which might train my thinking pattern") or having experienced its effectiveness in their life (3; e.g., "… when I faced the similar situation in life, the alternative opinion came to my mind and made me relief"). Nine individuals thought these types of exercises are ineffective, viewing them as self-deception (3; e.g., "… it likes usefulness self-deception …there was not so much kindness in life"), unable to solve real-life problems (2; e.g., "… it could not change my situation"), or presented in a boring manner (2; "…the idea is good but I dislike its presentation, so boring"). The remaining 11 participants believed the program may not be immediately effective but will have positive effects after a period of exercise, depending on length and frequency of practice (7). One of these conditional believers and other 2 participants (a total of 3) reported that certain scenario resolutions induced negative moods and resistance to being brainwashed, which needs time to overcome. The negative moods were from pushing them to positively interpret their familiar scenarios where they got hurt. Another one participant of those 7 participants believed the effectiveness in long-term context proposed his worries of negative relapse occurring again if trigger come. The rest 2

conditional effectiveness believers thought this program might be effective on those with mild symptoms.

Table 6. 12 Key Reasons (Themes) to Evaluation on eiIBM_RobotV1 Effectiveness

Evaluation	Key Reasons and subthemes	Examples
Effective (14)	Perceived logical validity and subconscious training (6)	
	Experiential benefits in life (3)	
	Engagement enhanced effectiveness (1)	
Ineffective (9)	Perceived lack of real-world kindness made it like self-deception (3)	
	Doubted positive thinking's problem-solving ability (2)	
	Boring presentation style (2)	
	Did not expect personal help (1)	
Conditional effectiveness (11)	Gradual effectiveness with sustained practice over time (7)	
	Scenarios' irrelevance triggered negative moods (3)	
	Concerns over possible relapse if triggered again (1)	
	Ineffective for serious problems, only helps mild symptoms (2)	
	Initial resistance, but improved after multiple exposures (1)	

Note. The number in brackets represents the number of people pick up corresponding evaluation.

The perceived effectiveness of the **eiIBM_RobotV1** appeared to depend largely on whether participants believed positive interpretation could address deep emotional issues. Those who felt the program logically trains subconscious thinking patterns were more positive. This indicated the first design improvement implication: explain the logical reasoning behind practices to convey cognitive skills developed rather than just promoting positivity (e.g., #148).

Experiencing benefits also increased trust in effectiveness (e.g., #165). Those seeing conditional effectiveness emphasized the need for sustained practice and gradual progress over time. (e.g., 7 participants emphasized the need for long-term practice). Framing emotional resilience to withstand triggers was also emphasized (e.g., #170). This suggests the second design improvement implication: position the **eiIBM** program as building emotional resilience tools rather than promising immediate transformation, establishing realistic expectations about gradual gains.

However, some doubted positive thinking alone could help with serious problems (#134). Others felt scenarios irrelevant to their experiences made the program seem like self-deception (#158 & #110). Boring presentation style further reduced engagement for some (#133 and #157). This points to the third design improvement implication: customize scenarios and interactions for personal relevance to counter feelings of self-deception from standardized content. Warm, empathetic embodiment can further bond and openness (#99).

6.4.2.5 Summary

The qualitative analysis addressed four interconnected questions to understand user perspectives on robots delivering **eiIBM_RobotV1**:

Qualitative question 1 – "What significant evaluations of the different robots emerged from the interviews?" Results identified seven key evaluation themes - concentration, companionship, oppression, interactivity, conviction, unnecessity, and unnaturalness, highlighting nuanced differences in how the telepresence, audio, and chatbot robots were perceived.

Qualitative question 2 – "Which factors shaped participants' preferences and avoidances for specific robots?" Results showed preferences aligned with strengths and deficiencies in the key evaluation areas. For example, the audio bot's concentration and companionship drove its selection, while the telepresence robot's physicality was pivotal for some preferring interactivity but distracting for others. Mismatched attributes and needs increased chatbot avoidance.

Qualitative question 3- "How do emotional distress triggers and expectations relate to robot selections?" Results indicated that robot preferences aligned with psychological needs stemming from triggers. Those with tangible triggers prioritized concentration from the audio

bot to address defined concerns, while participants with intangible triggers valued the telepresence robot's companionship for temporary relief without resolving trigger causes.

Qualitative question 4 – "What are participant-inspired insights for improving perceptions of effectiveness and engagement with the **eiIBM_RobotV1** program?". Results demonstrated the importance of improving the exercise rationales, framing the program as emotional resilience skills over time, and customizing content while leveraging robot relational capabilities.

6.5 Summary and Discussion

6.5.1 Discussion on Research Questions

This study conducted a within-subject experiment with three types of robots (audio robot, telepresence robot, and chatbot) as independent variables, followed by a semi-structure interview to explore the experience difference among robots delivering **eiIBM_RobotV1** and the reasons behind this difference. The study addressed two research questions:

RQ 2.1: Does similarity of experience occur within an individual interacting with different robots?

RQ 2.2: How do depressed individuals compare affordances to goals and concerns in terms of valence and relevance?

For **RQ 2.1**, the quantitative results showed no significant difference in most experiential variables across the three robot modalities, suggesting that depressed individuals perceived comparable experiences with the text chatbot, audio bot, and telepresence robot. However, participants reported a greater intention to use the audio robot compared to the text chatbot. This indicates some divergence in acceptance for different embodiments.

For **RQ 2.2**, the qualitative findings revealed how depressed individuals evaluate robotic affordances by comparing them to personal goals, needs and concerns stemming from

their emotional distress triggers and expectations. Key needs centered around concentration, companionship, and emotional support. Participants assessed the valence of each robot's affordances based on its perceived relevance for satisfying those needs. For example, the audio bot's concentration affordance elicited positive valence by aligning with the goal of focusing to address tangible distress triggers, while the telepresence robot's embodiment afforded companionship and interactivity that matched intangible triggered individuals' relief goals.

6.5.2 Theoretical and Practical Implications

The qualitative findings help explain the unexpected result from Study 1 that the chatbot elicited similar experiences as the audio bot and telepresence robot, despite different affordances in ease-of-use interaction. The within-subject comparisons revealed that some users likely preferred the audio bot to meet concentration needs, while others favored the telepresence robot for companionship. However, these two robots had some counter-good features. The chatbot had the least affordances and triggered fewer expectations to compare with.

This highlights the need to consider alignment between perceived and designed affordances as robot capabilities expand. With more features, user expectations also multiply, or it induces the unnecessity evaluation of the add-on features. For example, adding the visual modality into the robot, turning it into the telepresence robot, could give a sense of companionship for people's needs. The perceived and designed affordance synchronized, which might facilitate the positive comparison with the personal needs derived from emotional triggers and individual expectations. However, for those less needing companionship, the designed affordance of companionship was not perceived and was perceived as unnecessary compared to their emotional triggers and individual expectations.

This finding also helps to understand the moderating effect of **Medium** on the path from ease-of-use affordance to use intention of the robot through valence of ease-of-use in the present study. The relationship from ease-of-use affordance to ease-of-use valence was only significant with the audio bot and chatbot, but not with the telepresence robot. It might be because the ease-of-use affordance within the audio bot and chatbot was perceived and became the source of ease-of-use valence. However, the ease-of-use affordance within the telepresence robot was not a contributor to ease-of-use valence. The participants interacting with the telepresence robot perceived more features, and this caused more evaluations that harmed the minimal set of ease-of-use affordance transforming into positive valence.

The qualitative finding also implied that the ease-of-use content is different for individuals, according to their needs and emotional triggers. Ease-of-use encompassed concentration, companionship, interactivity - aligned with personal needs. In practice, researchers should give participants the choice to let them choose their personal robot assistant for **eiIBM** exercises, as there is no one type of robot superior to others.

Critically, emotional relief affordance depends partly on implementation factors beyond just the robot. Ensuring this affordance is positively perceived will be key for the next iteration. For example, high ease-of-use valence may not sufficiently drive use intentions if emotional relief valence is low. This likely explains the moderate robot use intention and program usefulness correlation in present study.

6.5.3 Design Implication

Study 2 provides insights to improve both the robot modalities (*eiIBM_Medium*) and the **eiIBM** exercise format (*eiIBM_Implementation*) for **eiIBM_RobotV2**.

Regarding the robot modalities, some participants avoided the telepresence robot because it reminded them it was not human, which harmed engagement. Some even felt

stressed or scared looking at the robot. Participants suggested preferring a human-like avatar over an obvious machine. Therefore, avatars can be utilized instead of physical robots. Avatars provide flexibility in appearance and behavior impossible with physical robots. This can help make interactions more natural and customized to the speaking content. Avatars also enable greater automation of the program in delivery thanks to advances in computer vision.

In addition, some features of the telepresence robot were seen as unnecessary, likely because they did not match individuals' needs and expectations. Therefore, robot appearance, facial expressions, and motions will be designed to clearly convey **eiIBM** content and emotional support. Matching features to their intended tasks remains critical, as irrelevant features may have detrimental effects. It emphasized task-contingency design (Hoorn & Huang, 2024).

Regarding **eiIBM** exercise format, explanations of the logical reasoning behind the positively resolved scenarios, expectation management about gradual gains, and relating content to personal experience can enhance trust on the program's effectiveness, as suggested by participants. Specifically, three improvements in **eiIBM_RobotV2** were identified: 1) providing additional attributions rationalizing positively resolved endings after affirming users' ideas, 2) using the agent to continually adjust user expectations about exercising the thinking pattern over time rather than persuading their belief on everything good in life, and 3) offering open-ended resolution for them to fill out instead of requiring them imagined the pre-designed positive outcome. Respecting multiple perspectives and supplementing scenarios with personalized relevance can satisfy desires for flexibility while reducing feelings of offense.

In the next study, I will improve **eiIBM_RobotV2** accordingly and re-analyze their experience and perception on the improved robot modalities versus **eiIBM** exercise format

and whether the intervention effect for negative interpretation bias and depressive severity remains.

6.5.4 Limitation of Current Study

Despite the implication in theory building and robot therapy implementation, the study design and sample characteristics do present some limitations that should be addressed in future research. The within-subject design inherently carries potential carryover effects, where experiences with earlier robot modalities may influence perceptions of subsequent ones. While the author attempted to mitigate this through randomization of exposure order, some residual carryover effects may still be present, as participants completed the questionnaires immediately after each interaction rather than providing a summative evaluation. With the relatively small sample size, the author was unable to fully examine the impact of exposure order.

Additionally, the gender imbalance in the sample, with 27 out of 40 participants being female, may limit the generalizability of the findings. Though gender was not correlated with the key variables, it is possible that gender could interact with the order of robot modality exposure to influence individuals' preferences and experiences. For example, the order effect may manifest differently for male and female participants, potentially due to gender-based differences in information processing or decision-making tendencies.

The theme coding with qualitative data was conducted solely by the author, raising the possibility of unintentional biases in the interpretation of the interview responses. Employing multiple coders and assessing inter-rater reliability could have enhanced the reliability of the qualitative findings. Due to resource and time constraints, the author was unable to incorporate this approach in the current study. However, for future research, the

author plans to involve additional coders and implement measures to ensure the objectivity of the qualitative analysis.

Chapter 7: Effect of eiIBM_RobotV2 (Study 3)

This chapter explores the experiences elicited by two types of robots—a virtual avatar and an audio bot—delivering an improved online intervention for Imagery-enhanced Elaborative Interpretation Bias Modification (**eiIBM_RobotV2**). Building on the findings from Study 1 and Study 2, **eiIBM_RobotV2** incorporates improvements to the robot types and **eiIBM** exercise format. Artificial intelligence is used to increase autonomy, self-relevance, and naturalness of the interactions. By examining experiential variables derived from the I-PEFiC model across time and their effect on cognitive outcomes (BDI-II, WSAP-D, SST, SRT), this chapter empirically investigates the influence of robot type on therapy over time.

The results are also compared with those from Study 1 to better understand the effect of **eiIBM** exercise format (*eiIBM_Implementation*) and robot (*eiIBM_Medium*) features on user experience and intervention outcomes. The findings aim to shed light on how users' perceptions and emotional distance towards robots change with sustained exposure in the context of robot-delivered therapy. This contributes to the understanding of incorporating social robots into AI-enhanced, empirically evidenced therapy.

7.1 Introduction

Study 1 revealed no significant differences in user experiences between groups guided by chatbots, audio bots, and telepresence robots for **eiIBM_RobotV1**. However, emotional distress relief valence moderately contributed to **eiIBM_RobotV1** use intention, while ease-of-use interaction valence and relevance did not, suggesting the importance of emotional distress relief affordance for willingness to use **eiIBM_RobotV1**. Overall, incorporating robots benefited negative bias reduction and depressive symptoms.

Study 2 showed that the lack of differences in user experiences across robot embodiments was largely due to the counteracting effects of certain robot features and a mismatch between individuals' needs and the capabilities offered by each robot. Participants who preferred audio bots valued the concentration and companionship provided during the exercise, while those who desired greater interactivity found the telepresence robot pivotal. However, for some participants, the physical embodiment of the telepresence robot was visually distracting.

Based on the findings from Studies 1 and 2, Study 3 aims to improve the delivery of **eiIBM** by better aligning both the robot modalities (*eiIBM_Medium*) and **eiIBM** exercise format (*eiIBM_Implementation*) with user needs identified in the previous studies. Key changes include:

1. Replacing the narrow ending resolution provided in the exercise scenarios with a more open-ended resolution to improve participants' autonomy and self-relevance of the scenarios (changes in *eiIBM_Implementation*).
2. Using virtual avatar instead of physical telepresence robots to deliver the exercise instructions and guidance, preventing perceptions of stiffness, oppression, or feeling of eeriness (changes in *eiIBM_Medium*).
3. Aligning the avatars' features, behaviors, and responses more naturally to the cognitive restricting task in order to enhance ease-of-use perceptions (changes in *Medium_Behavior*).

The present study re-evaluates the effect of this improved robot-delivered intervention (**eiIBM_RobotV2**) on both user experience and intervention outcomes, addressing three main research questions:

RQ3.1 Are user perceptions largely similar between virtual avatars and audio bots delivering the **eiIBM_RobotV2** exercise?

RQ3.2 Given the improved **eiIBM_RobotV2** exercise, do the intervention outcomes differ depending on whether it is delivered by virtual avatars or audio bots?

RQ3.3 Dose **eiIBM_RobotV2** produce higher or at least comparable intervention effects on negative interpretation biases and depressive symptoms compared to **eiIBM_RobotV1**?

Study 2 showed that audio bots provided the needed concentration and companionship that many participants valued during the exercise. Virtual avatars can counter some limitations found with telepresence robots by offering more flexible and customizable behaviors and features while avoiding the potentially oppressive physical presence. By aligning the avatar's features closely with the cognitive exercise, it is expected that both virtual avatars and audio bots will elicit largely equal perceptions of ease-of-use affordance (**H6.1**) that positively transforms into similar valance and relevance comparison (**H6.2**), leading to comparable robot use intentions across both embodiments (**H6.3**). Additionally, the vivid facial expressions possible with virtual avatars may increase participants' evaluations of the program's usefulness and their overall trust in it (**H6.4**) compared to audio bots.

With sustained exposure to **eiIBM_RobotV2**, previous experience should shape subsequent experience. Prior responses do not linearly influence new comparisons but establish thresholds where good experiences enable fair future assessments and bad experiences worsen future assessments (Stafford et al., 2010). Since **eiIBM_RobotV2** meets emotional relief and ease-of-use interaction needs, it is hypothesized that with sustained exposure, the experience of previous interactions will contribute to equal perceived relevance and valence of ease-of-use affordance from the avatar and audio bot in subsequent interactions (**H6.5**).

Regarding **RQ3.2**, equal intervention effectiveness is hypothesized between virtual avatars and audio bots (**H6.6**), given that the core **eiIBM** exercise mechanism remains the same across both conditions. As for **RQ3.3,** it is expected that **eiIBM_RobotV2** produce

higher intervention effects for those with mild to moderate levels of depression compared to **eiIBM_RobotV1** (**H6.7**), since the changes made improve personal autonomy and relevance, which matches the needs expressed by some participants in Study 2. However, those with more severe depression may show comparable or even lower intervention effects with **eiIBM_RobotV2** (**H6.8**) due to the increased difficulty and cognitive load of generating their own open-ended resolutions.

To test these hypotheses, a between-subjects experiment was conducted comparing virtual avatars and audio bots delivering the improved **eiIBM_RobotV2** exercise. Before the testing, the design of **eiIBM_RobotV2** will be introduced.

7.2 Design Strategy of eiIBM_RobotV2

7.2.1 Automatic Response System in eiIBM_RobotV2

The automatic response system in **eiIBM_RobotV2** is designed to resolve ambiguous scenarios by generating responses that judge the coherence and sentiment of the user's input in relation to the given context. Coherence, which refers to the local consistency among sentences, is used as an indicator of successful ambiguity resolution, as it is subjective and hard to define. Sentiment polarity is another metric for evaluating the resolution, as resolved scenarios tend to have a clear sentiment.

Coherence modeling has been a long-standing topic in discourse analysis (Lapata, 2003). While sentence ordering tasks are commonly used to evaluate coherence models in Natural Language Processing (NLP), their limitations have been noted, as high performance on these proxy tasks does not necessarily indicate the ability to identify order-insensitive text (Lai & Tetreault, 2018). To address this, the **eiIBM_RobotV2** system incorporates common-sense reasoning tasks, which require the model to go beyond pattern recognition and make inferences using world knowledge. Datasets like Even T2Mind (Rashkin et al., 2018) and

SWAG (Situations with Adversarial Generations, Zellers et al., 2018) are particularly relevant to the ambiguity resolution task, as they focus on reasoning about the intents and reactions of participants in a given event.

Sentiment classification is another key component of the automatic response system, enabling it to analyze the opinions and attitudes expressed in the user's input (Wen et al., 2020; Birjali et al., 2021). By integrating datasets such as CMU-MOSEI (Zadeh, 2018) and SST-2 (Socher et al., 2013), the model learns to associate text inputs with sentiment labels, which is crucial for determining the emotional tone of the resolved scenario.

The core functionality of the automatic response system can be divided into two target tasks: classification and generation. The classification task involves determining the coherence and sentiment of the user's input in relation to the given context, with the output being one of four categories: positively coherent, negatively coherent, positively incoherent, or negatively incoherent. Based on this classification, the generation task produces appropriate feedback to guide the user through the ambiguity resolution process.

To implement this system efficiently, **eiIBM_RobotV2** leverages pre-trained language models such as GPT, BERT, and Llama, which can be fine-tuned for various natural language understanding tasks such as reading comprehension (Radford et al., 2018) and natural language inference (NLI) (Devlin et al., 2018). The stack of bidirectional transformer encoders in these models enables transfer learning, allowing the system to build upon the knowledge gained from large-scale pretraining and adapt it to the specific requirements of ambiguity resolution.

While pre-trained language models have limitations in terms of explainability and interpretability (Zhao et al., 2024), they provide a practical solution for rapidly deploying a conversational AI system for mental health applications. By fine-tuning these models on small, tailored datasets relevant to the **eiIBM** exercise, the automatic response system can

generate context-sensitive responses that effectively guide users through the ambiguity resolution process.

7.2.2 eiIBM_RobotV2 Task

The **eiIBM_RobotV2** consists of three key subtasks: casual chatting and greeting, the **eiIBM** exercise itself, and concluding with feedback (Figure 7.1). The automatic response system is embedded into the **eiIBM** exercise (Subtask 2).

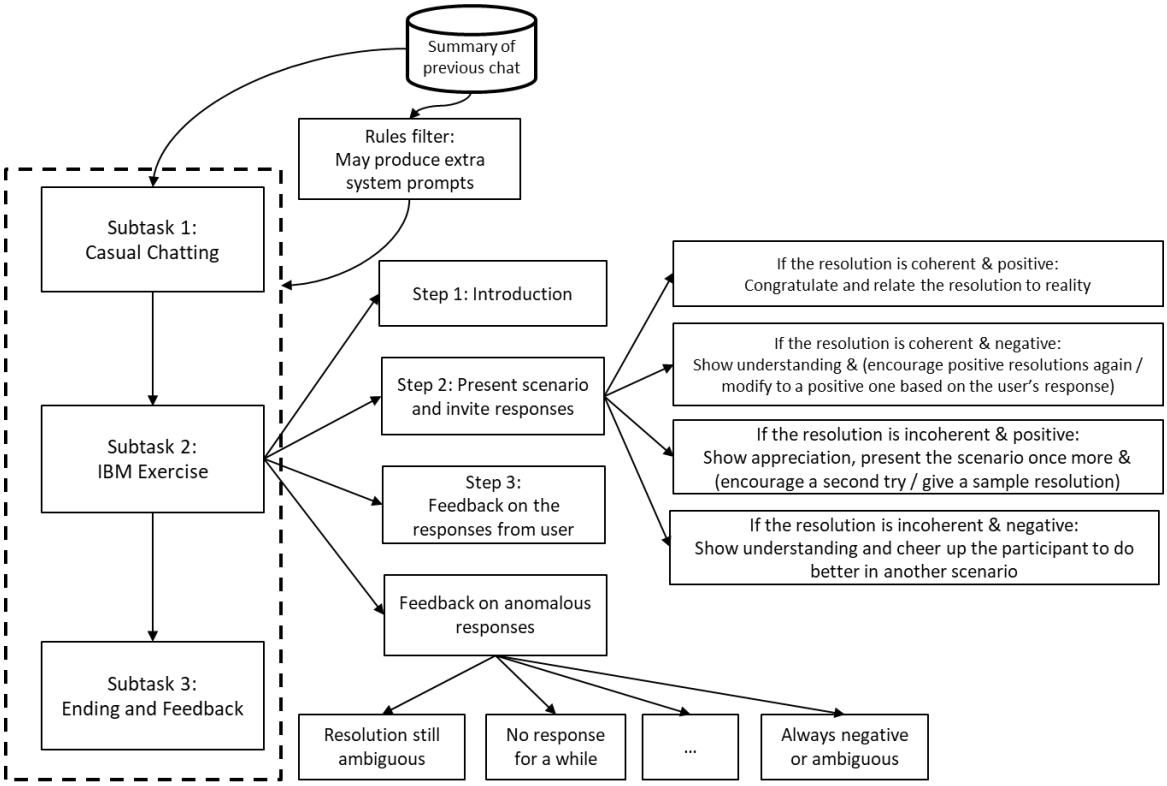

Figure 7.1 Task Process of the eiIBM_RobotV2 Exercise

7.2.2.1 Subtask 1: Greeting/Casual Chatting

In the first subtask, the robotic agent greets the user and engages in casual conversation, inviting the user to share their daily life experiences as a warm-up. The agent responds empathetically, establishing user engagement. To provide a sense of continuity and personalization, the agent considers previous chat summaries and addresses the user by name.

The quality of this interaction is crucial, as it sets the tone for the rest of the exercise and plays a significant role in building friendship.

7.2.2.2 Subtask 2: eiIBM_RobotV2 Exercise

The second subtask, the **eiIBM_RobotV2** exercise, consists of three steps:

1. Introduction: The robotic agent invites the user to start the exercise, prompting them to prepare mentally. In the first encounter, the agent briefly explains the task and addresses any questions.
2. Presenting scenarios and inviting responses: The agent presents various scenarios to the user one by one, inviting positive interpretations or denouements from the user's responses.
3. Providing feedback: After the user speaks, the agent offers specific feedback based on the user's response. For coherent and positive responses, the system congratulates the user and relates the resolution to reality. For coherent but negative responses, it demonstrates understanding and encourages more positive resolutions. If the resolution is incoherent but positive, the system appreciates the effort, re-presents the scenario, and encourages a second try or provides a sample resolution. For incoherent and negative responses, the system shows understanding, motivates the user, and suggests improved responses.

The feedback's delivery is of paramount importance, as the exercise's effectiveness is substantially influenced by the user's experience (according to the findings in Study 1). An effective response often possesses an empathetic tone, is framed optimistically and upliftingly, and is concise, precise, and clear. By connecting the user's response to the daily-life context, the robotic agent's response should also be encouraging, appreciative, engaging, insightful, non-judgmental, and patient.

7.2.2.3 Subtask 3: Ending and Feedback

The final subtask concludes the session and provides overall feedback. The robotic agent appraises the user's effort, addresses their needs or concerns in daily life (especially from their sharing during the greeting), and looks forward to the next meet-up. In the last wave, the agent also asks for feedback on the user experience. This stage is critical for reinforcing the lessons learned during the exercise and preparing the user for future waves.

7.2.3 Controlling ChatGPT 3.5 turbo for eiIBM_RobotV2

The generative AI of large language models (LLMs), specifically GPT 3.5 turbo (or ChatGPT), was incorporated into the **eiIBM_RobotV2** program. Although other models like Llama and Claude2 became available in July 2023, offering benefits such as increased token capacity and open-source access for researchers, my prototype was built using ChatGPT, the state-of-the-art model at the start of prototyping in June 2023.

While ChatGPT cannot automatically deliver therapy, controlling methods like prompt engineering, fine-tuning, and reinforcement learning with human feedback (RLHF) can coax pretrained language models to perform AI tasks (Stade et al., 2023). Prompting involves casting a task as a textual instruction to a language model (Liu et al., 2023), either manually crafted or automatically generated using fill-in templates for token, span, and sentence-level completion (Petroni et al., 2019; Brown et al., 2020; Shin et al., 2020). This makes prompting applicable to more challenging NLP tasks, such as QA, MT, and summarization (Schick & Schütze, 2021). Prompts can be zero-shot, providing context and framing without examples, or few-shot, including examples (Stade et al., 2023).

Fine-tuning allows customizing the interaction scenario at low cost by changing parameters on the top transformer layers of the LLM using a limited set of human-labeled input-output data (Stade et al., 2023). OpenAI announced the fine-tuning function of

ChatGPT for me to fine-tune the model with small datasets. Reinforcement learning from human (RLHF) is an advanced supervised fine-tuning method that collects qualified prompt-completion examples with human annotations, trains a reward model to guide the LLM, and incorporates real-time human preference ratings (Stade et al., 2023). However, RLHF was not considered due to the lack of low-level access to model training (not feasible for OpenAI API users) and sufficient conversational data for annotation.

Considering the viewpoint of Stade et al. (2023) on integrating clinical language models into psychotherapy using collaborative AI ("human in the loop") in early stages before the AI system proved safe for deployment in behavioral health, prompt engineering was chosen as the most suitable method for timely controlling ChatGPT to well deliver the **eiIBM** exercise. Previous study (Fairburn & Patel, 2017) on collaborative language models may involve a model delivering a semi-independent structured intervention like a chatbot, with a provider monitoring and taking control as needed, similar to guided self-help. Fine-tuning was abandoned due to its excessive cost and limited effect.

Although prompting eliminates the need for fine-tuning, identifying good prompts can be challenging (Liu et al., 2023). ChatGPT's performance has shown some limitations with prompt due to the complexity of the task (e.g., the standardized intervention – **eiIBM** exercise) and the definite memory, leading to deviation of the general or earlier-mentioned instructions (e.g., the system prompt) and unexpected behavior, such as lacking empathy or providing irrelevant responses. These shortcomings are intrinsic to the current state of generative AI.

To overcome these limitations and address LLM's low explainability (e.g., interpretability; Angelov et al., 2021) and vulnerability to irrelevant user prompts (e.g., easily interrupted by the user's irrelevant user prompt), a conversation monitor was designed to

generate live prompts guiding the **eiIBM** exercise, and intermediate filtering layers were added to inspect input and output.

7.2.4 Design Strategy of the Experimental Version of eiIBM_RobotV2

Figure 7.2 illustrates the designed architecture of the **eiIBM_RobotV2**. These filters sandwich the central task monitor (i.e., conversation monitor), applied before the input text is processed by the ChatGPT algorithm or before the response is delivered to the user. The speech-to-text API converts user's response to the input of the input filter. The input filter analyzes, purifies, and rephrases this input to suit the task process. Exceptionally long responses may be summarized, influent and rude expressions are rephrased to be neutral and linguistically understandable. The response towards the scenario is organized in the way that projects the information that the task process needs to analyze.

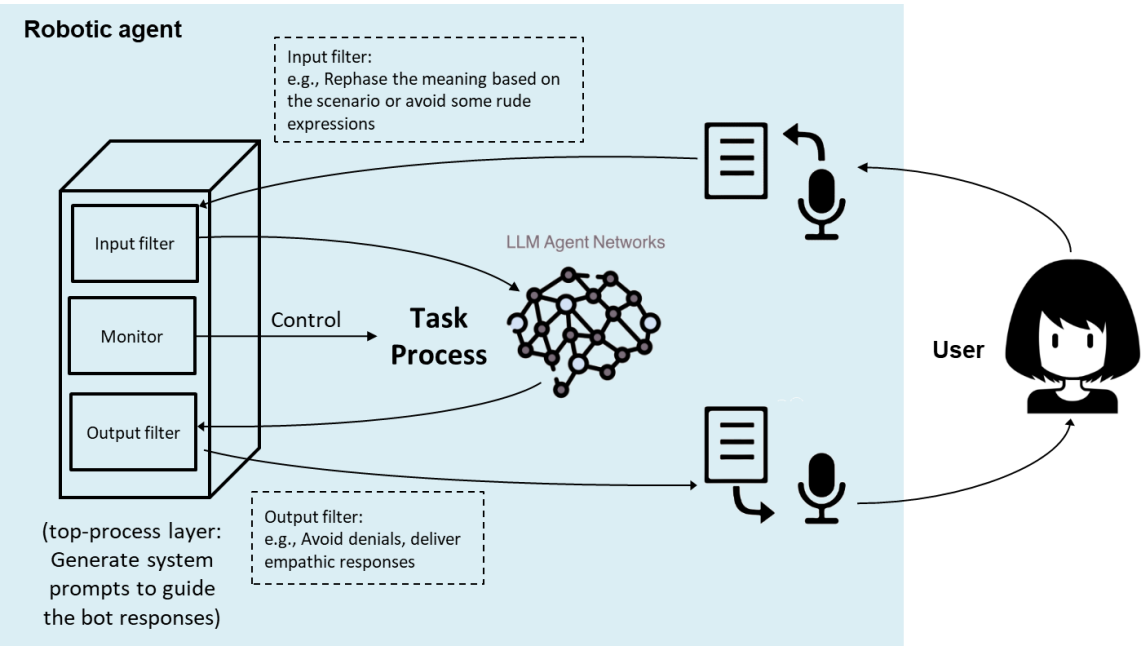

Figure 7.2 Information flow of the additional layer between the user and the robotic agent

This input filter can be adapted throughout the task. For *Subtask 1*, a rules filter could generate additional personalized system prompts based on previous chat content, such as the user's name or past experiences to better engage the user. For *Subtask 2*, the intermediate

input filter could provide nuanced handling of various user responses. The user's response towards the interpretation or denouement of the scenario may be scattered and unorganized, leading to inaccurate analysis, or even interfering with the nominal task, producing unexpected output. The input filter can reorganize the input to be a task-process–friendly format. Moreover, the input filter could detect and handle anomalous responses, such as ambiguous or persistently negative feedback and prolonged idle. This allows the task process to produce responses that are not within the main tasks temporarily, enhancing the system's adaptability and responsiveness (Figure 7.2).

The output filter focuses on polishing the manner of response delivery to the user. To produce responses as described in the subtasks, the output filter transforms denials to encouragement, judgement (of unsatisfactory response) to guidance, and written language to spoken language, insert the name of the user, add variations of delivery to be more human, maintain an empathic tone, emphasize appreciation, praise and affirmation, and so on. The polished response would enhance human–robot interaction, allowing better engagement from the user, providing an environment with more positive feedback to motivate the user to be more attentive, pay more effort in the exercise, and make practice more regularly. The exercise can then alter the user's rumination more effectively.

The criteria of the input and output filters and the quality of filtered text is monitored in this top-process layer. In other words, this layer can be regarded as the moderator of the quality control of the task process. Such functionality could be a rule engine such as a rule-based and inference engine or a small specific model through reinforcement learning. This approach enables better inputs for the task process and enhances the quality of the responses, improving the overall user experience.

7.3 Methods

7.3.1 Participant

The inclusion and exclusion criteria for Study 3 participants were identical to those in Study 1 and Study 2, with the exception that all participants were Cantonese speakers. This language criterion was implemented because language was found to correlate with experience in Study 1. To better control this effect, the study targeted only Cantonese speakers. Participants were recruited between August 2023 and October 2023 using the same methods as in Study 1 and Study 2. Eligible participants were asked to sign a consent form after reading an information sheet that outlined the study details (e.g., aims, length, involvement, randomization, incentives). In total, 44 depressed Hong Kong residents who spoke Cantonese (M_{age}= 24.57, SD_{age}= 3.91, 34 Female) completed the experiments. Participants who completed the experiment received HK$350 ParkShop cash coupons as compensation and were entered into a lucky draw (15% chance) for a Hanson's Professor Einstein™ robot. Additionally, they received an assessment report and the privilege to use the **eiIBM_RobotV2** program once it was launched.

7.2.2 Design

The study utilized an identical design to that in Study 1, using a between-subject (**Medium**: *Avatar* versus *Audio*) repeated (**Time**: *T1*, *T2*, and *T3*) design with a pre-(*pretest*) and post-assessment (*posttest*). Participants were randomly assigned to either the *Avatar* or *Audio* condition, with age, gender, and depression severity controlled through pre-assessment. The *Avatar* condition featured a virtual figure face on the screen (Figure 7.3), while the *Audio* condition did not (Figure 7.4); both conditions shared the same voice. A complete experiment was conducted over a 3-week period, involving two assessments and six interaction sessions with the robots, with a minimum 3-day break between sessions. In each interaction session, participants came to the lab, and the robots guided them through the

eiIBM_RobotV2 exercise. Participants completed a questionnaire after the 1st, 3rd, and 5th sessions to record their interaction experience. Cognitive assessments were completed before and after the six-session **eiIBM_RobotV2** program, with the pre-assessment done on the participants' mobile app **IBMTest@POLYUSD** at home and the post-assessment conducted at the lab.

Figure 7.3 Settings of virtual avatar guiding the eiIBM_Robotv2.

Figure 7.4 Settings of audio bot guiding the eiIBM_Robotv2.

7.2.3 Procedure

After successful registration, researchers scheduled a three-week experiment period with the participants. The day before the scheduled period, participants were sent links to

install the **IBMTest@POLYUSD** mobile app and complete the pre-assessments on the app. Participants were required to complete pre-assessment tasks sequentially without stopping within each task but could rest between tasks. They were reminded that unreasonable response times or inconsistent responses might lead to suspension of participation. After checking the pre-assessment quality, researchers scheduled the first exercise session with successful participants within the next 2 days. The remaining five sessions were scheduled either week-by-week or all at once, with at least 3 days between sessions and a maximum of 2 sessions per week. Participants were assigned to either the *Avatar* or *Audio* condition, but the allocation was concealed until their first lab session. Researchers sent reminders to participants about the next day's scheduled session.

At the scheduled time, researchers welcomed participants and guided them to sit in front of the computer. In the first session, a brief introduction and important points were provided. Once participants understood that they should engage in the exercise by speaking to the robot and not work on other issues during the exercise, researchers left the experiment room for an inner control room. During the exercise, the participants followed the robots to complete the **eiIBM_RobotV2** session. The interaction procedure is detailed in **7.2.4 Apparatus and Materials**. After the 1st, 3rd and 5th session exercises, the participants were guided to fill out a questionnaire inquiring about the perception and experience of the interaction. Researchers would thank the participants and let them leave after checking the questionnaire quality or directly if there was no questionnaire that day.

Upon completing all **eiIBM_RobotV2** sessions, participants completed the post-assessments via **IBMTest@POLYUSD** at the lab, following the same requirements and task sequence as the pre-assessment. Researchers checked the quality of the assessments and confirmed participants' completion of the entire experiment. Completed participants received HK$350 ParkShop cash coupons as compensation. After all participants finished the

experiments, researchers hosted the lucky draw via Facebook Live and arranged prize pick-up with the winner. Assessment Reports were delivered to participants via WhatsApp. As part of the participant incentives, the author had promised to provide access to the **eiIBM_RobotV2** platform upon its launch. However, the development of the platform was still ongoing at the time of thesis submission, and the author was unable to fulfill that promise by the deadline.

7.2.4 Apparatus and Materials

7.2.4.1 Pre- and Post-assessment Apparatus

IBMTest@POLYUSD and **IBMTestManagement@POLYUSD** were used to measure and manage the data collection on participants negative interpretation bias and depression severity, as described in Study 1.

7.2.4.2 eiIBM exercise Materials

7.2.4.2.1 Avatar

The avatar used in the present study was presented as a gender-neutral, child-like character projected on the computer screen facing participants. Rather than developing a fully automated virtual agent, a cost-effective approach was adopted using an iPhone15 Memoji avatar controlled through real-time Wizard-of-Oz style puppeteering. This involved the researcher animating the avatar's lip movements, facial expressions, and gestures coordinated with the lively generated verbal responses voiced by the audio bot. The resulting avatar animation was mirrored to the projection screen in real time Air Screen Mirroring Receiver. To ensure that only the avatar was visible and other areas, such as the iPhone15 notification bar, were covered, a window form with a transparent middle area was created on the mirroring image. The effect can be seen in Figure 7.3.

This puppeteering method allowed for the conveyance of coordinated nonverbal behaviors aligned with the avatar's speech, enhancing perceptions of natural interaction. Using a generic, anthropomorphic avatar avoided potential biases associated with more gendered or realistic human likenesses (McDonnell et al., 2012). While fully automated virtual agents with coordinated speech and animation are an active area of artificial intelligence research, current consumer-grade solutions cannot yet achieve the consistency and contextual relevance needed for the avatar in this study. Thus, the Wizard-of-Oz approach balanced conveying naturalistic cues while maintaining control over the avatar's behaviors to support the experimental protocol.

7.2.4.2.2 Interpretation Bias Modification Task Stimuli

Study 3 utilized task stimulus described in Study 1. The **eiIBM_RobotV2** improved with an open-ended resolution of the ambiguous scenarios rather than the purely judging narrow-ended fill-ins. This improvement made each round of the **eiIBM_RobotV2** exercise longer. To control the exercise length, only 13 out of the 21 ambiguous scenarios in each session.

The scenarios were transformed into the format required for the **eiIBM_RobotV2** exercise. Each round of the exercise consisted of an ambiguous scenario without resolution and a corresponding image with a descriptive title. A sample adapted scenario is as follows:

[image with title] A image with the scenario title

[transcripts of describing the scenario] You walk on the street; you see your neighbor not far away from you. You call him by name and try to greet him, but he doesn't answer you. You may be thinking it is because _____. (你走在擁擠的大街上，你看見你的鄰居在離你不遠處。你叫他的名字試圖跟他打招呼，但是他沒有回應你。你想這可能是因為_____)

7.2.4.3 eiIBM_RobotV2 Operation Setting

In the experiment, the participant is seated in front of a "computer" in the participant room and asked to initiate an Internet connection to a robotic agent. The robotic agent engages in an oral conversation with the user, while the screen presents a series of scenarios with or without a live avatar. Due to technical constraints, the visual and audio elements of the participant's experience had to be launched in separate systems, necessitating signal synchronization. To mimic such an interface in the participant room, a pseudo-synchronous configuration was adapted in the Wizard-of-Oz (Woz) setup, hard-wiring both elements from the operation room to the participant room (Figure 7.5).

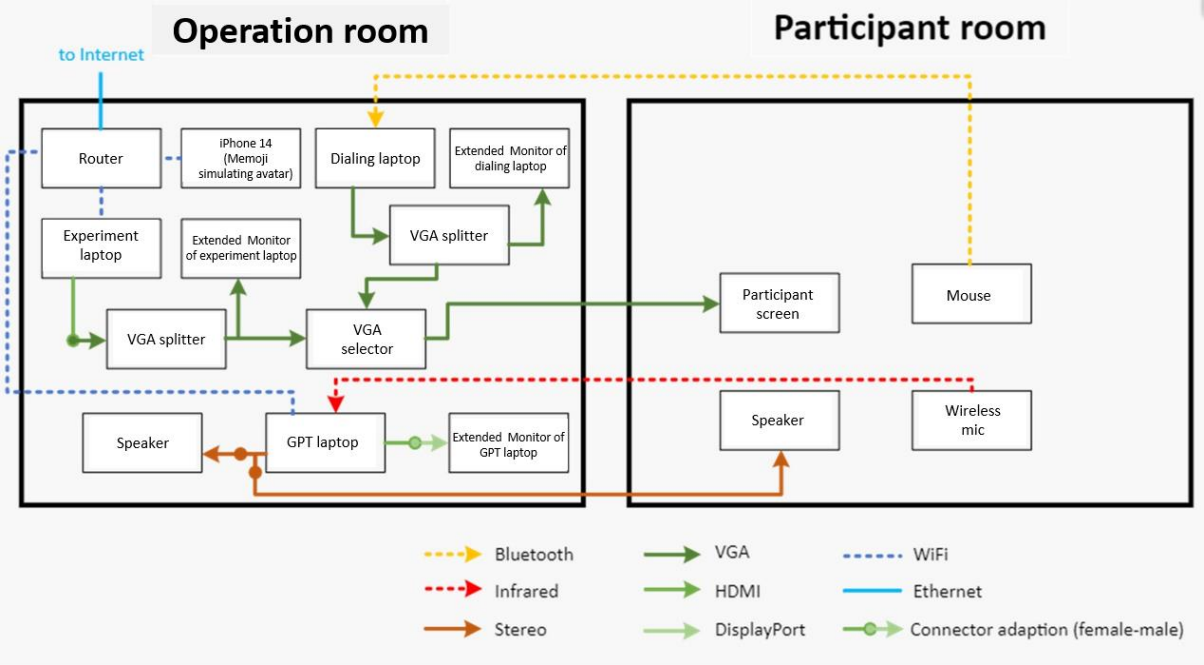

Figure 7. 5 Devices and their connections for Wizard-of Oz

The pseudo-synchronous configuration addresses the inconsistent latency of signals coming from separate systems over the network. The "computer" in the participant room is actually a screen (Participant Screen) connected to the operation room, where wireless I/O devices control the Dialing laptop. The Internet connection is emulated by the Dialing program (Figure 7.6) on the Dialing laptop. After the participant clicks to start the

connection, the operator in the control room switches the display of the Participant Screen from the Dialing laptop to the Experiment laptop, which has the conversation interface (Audio/Robot experiment program, see Figure 7.3 and Figure 7.4) open. VGA display was used due to its shorter switching time compared to HDMI display.

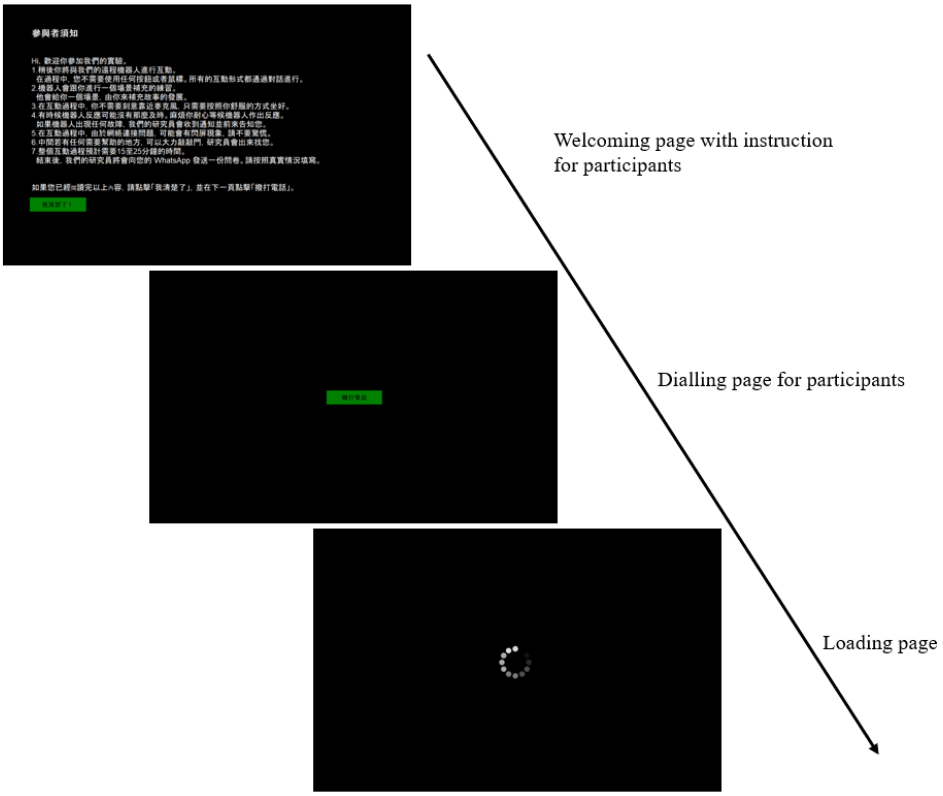

Figure 7. 6 Dialing program interface in Study 3

In the *Avatar* condition, an actor in the control room controls the facial expressions of the memoji to emulate talking, listening, etc. The memoji is mirrored from an iPhone with Face ID using Air Screen Mirroring Receiver installed on the Experiment laptop, overlaid with the Robot experiment program. The participant's audio input is fed into the speech-to-text API in Google Chrome on the Audio laptop, which becomes the input text for the GPT interface program in GPT laptop. The operator makes necessary clarifications for the participant input, such as correcting wrongly identified or missing text. Additionally, the operator can select specific prompts from the Experiment prompt utility to the

eiIBM_RobotV2 operation system (GPT interface program) as additional system prompts for further instruction.

The automatic response system inside the **eiIBM_RobotV2** operation system produces the response from the input text and the output text is converted to speech using the text-to-speech API in Google Chrome which is then fed to the speaker in the participant room.

To protect the participants privacy while utilizing large language models, the author had several safeguards in place. On the one hand, all interactions were closely supervised by the research team. They received and confirmed the automatically generated responses before sending to ensure appropriateness and avoid unsuitable content. As described in P236-P237, the author also pre-programmed prompts to guide the robot on improving responses if needed. On the other hand, communication was carefully scoped to avoid inquiring about private or sensitive information, focusing solely on the **eiIBM** exercise without delving into emotional or personal topics. During the entire experiment, the participant was isolated from the control room with two doors in series and thus did not have any knowledge of the experimental configuration.

The complete layout of the experimental setting without the connection of the devices in Study 3 is shown in Figure 7.7.

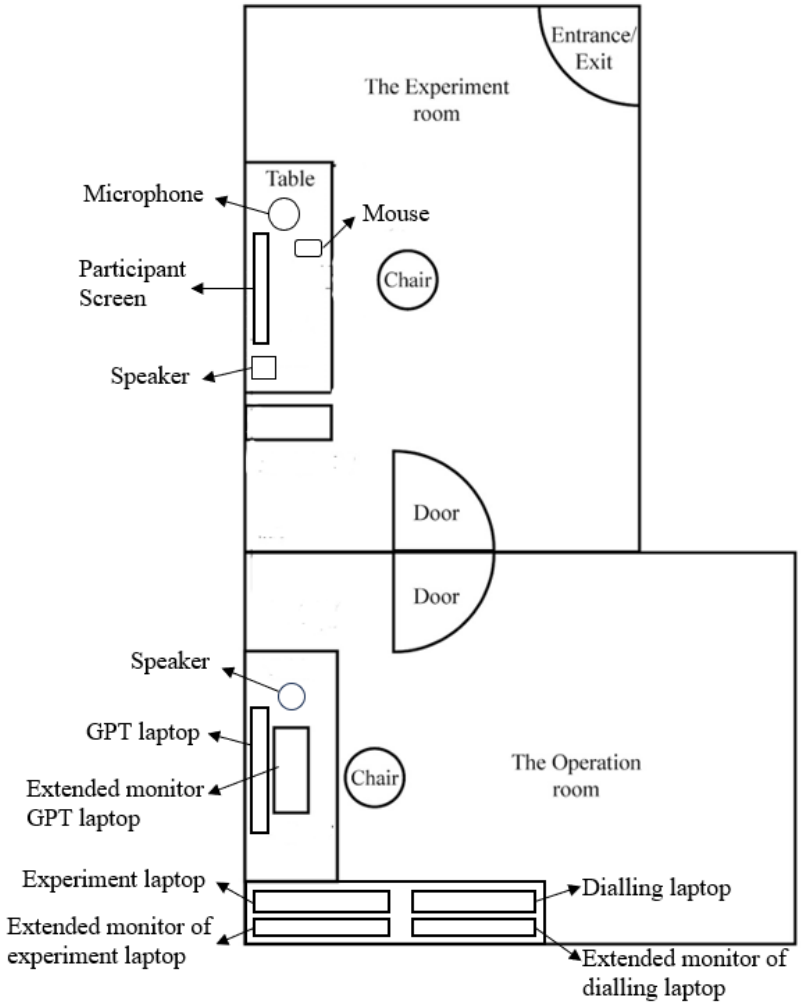

Figure 7.7 The experimental setting without the connection of the devices in Study 3.

7.2.4 Measures

7.2.4.1 Cognitive measurements

The cognitive measurements used in Study 3 were the same as those in the preparation study, including *DS_MS*, *SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*.

7.2.4.2 Experiential measurements

In Study 3, the experiential variables were measured at three timepoints during the experience, similar to Study 1. These variables were derived from I-PEFiC, focusing on the affordance of ease-of-use interactions (See Figure 7.6). The predictor of relevance and valence of ease-of-use affordance was prior experience on the program (*Pre_Exp*). In the first session, *Pre-Exp* was based on the perceived ease-of-use affordance, while in the subsequent sessions, the experience of usefulness and trust on the program dominated the ease-of-use affordance as the predictor.

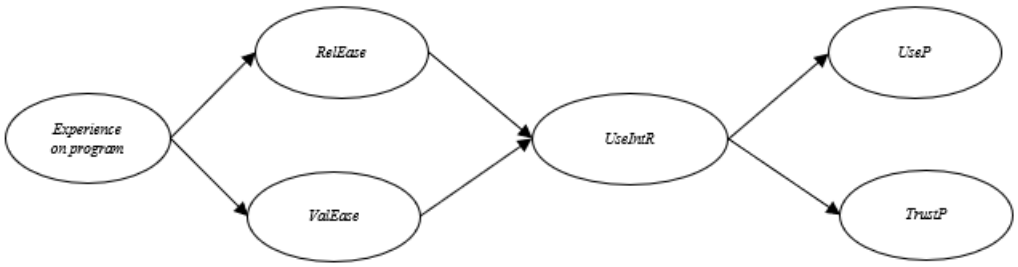

Figure 7. 8 The relationships of the experiential variables derived from I-PEFiC for Study 3

Study 1 revealed that in **eiIBM_RobotV1**, the affordance of intervention delivery was perceived, and the valence of such affordance contributed to the intention to use the program. The intervention outcomes also demonstrated the effectiveness of **eiIBM_RobotV1**. Based on the findings from Study 2, Study 3 improved the **eiIBM** exercise format (*eiIBM_Implementation*) by allowing the participants to self-generate positive resolutions for the scenarios in the exercises. Accordingly, Study 3 enhanced the robot modalities (*eiIBM_Medium*) and their behavior (*eiIBM_Delivery*).

Believing that the improvements to the **eiIBM** exercise format did not negatively impact the perceived affordance of intervention delivery and its comparison to emotional relief goals, I focused on the ease-of-use affordance-related variables and their relationships.

This provided more insights into the incorporation of robots into CBT therapy for depressed young adults.

Therefore, I measured the affordance of ease-of-use interaction (*AffEase*), along with its relevance (*RelEase*) and valence (*ValEmo*) regarding the concern of ease-of-use interaction (*GoEase*), exploring robots' facilitating influence. Unlike in Study 1, which measured the use intention of **eiIBM_RobotV1**, Study 3 focused on the robot use intention. This decision was made because Study 2 showed that the program use intention might be affected by emotional triggers and personal needs. By focusing on the robot use intention, I aimed to counterbalance these affective factors and facilitate a positive evaluation of the entire program. Trust in the program and usefulness of the whole program were also evaluated to understand the relationship between the evaluation of the robots and the evaluation on the program. Additionally, I asked the participants about their evaluation criteria for trust (*TrustBase#*). In the second and third measure time sessions, I did not measure ease-of-use affordance anymore but merged the previous usefulness and trust on program as prior experience (*Pre_Exp*). This was inspired by the I-PEFiC model itself having bidirectional relationships between the constructs. Specifically, the end-evaluation state affects a new round of comparison on features and goals.

Several 10-point rating scales were used for manipulations checks on the ease (*Check_EaseImg#*) and engagement (*Check_EngImg#*) of mental imagery, overall exercise engagement (*Check_EngExcer#*), robot satisfaction (*Check_SatR#)*, tolerance to the robot false-making (*Check_TolR#*) and the perceived frequency of robot error (*Check_ErrFrequent#*). All the variables are shown in Table 7.1.

Table 7. 1 Items for Experiential Variables in Study 3

Construct	Categories	Experience source

Affordance	Affordance – Ease-of-use (*AffEase*)	Robot	*AffEase#_1*: Based on my experiences, the training program's interactions are clear and easy to understand *AffEase#_2*: Based on my experiences, I find that such training programs are easy to use *AffEase#_3*: Based on my experiences just now, I find that I can easily become proficient in the interaction process *AffEase#_4R*: Based on my experience just now, - it would take me a long time to get used to such a training program (R) *AffEase#_5*: Based on my experiences, I immediately understand how I should interact with the training program *AffEase#_6R*: Based on my experiences, the training program is a little difficult to use (R)
Relevance	Relevance – robot relevant to the goal of completing the exercise (*RelEase*)	Robot	*RelEase#_1*: In terms of providing help, robot - makes me complete tasks faster *RelEase#_2*: In terms of providing help, robot - improves my efficiency *RelEase#_3*: In terms of providing help, robot - helped me get involved in the training *RelEase#_4*: In terms of providing help, robot - enhances the training effect *RelEase#_5R*: In terms of providing help, robot - increased the difficulty of my training (R)
Valence	Valence – expectation on the comfortable experience (*ValEase*)	Robot	*ValEase#_*1: The experience with [name of robot] (guide), I felt – it is pleasant *ValEase#_2*: The experience with [name of robot] (guide), I felt - it is interesting *ValEase#_3R*: The experience with [name of robot] (the guide), I felt - is uninteresting (R) *ValEase#_4R*: The experience with [name of robot] (the guide), I felt - it is disappointed (R) *ValEase#_5*: In the experience with [name of robot] (guide), I felt – is happy
Use Intention	Use intention – robot (*UseIntR*)	Robot	*UseIntR#_1*: Based on my experience just now, I am willing to have this robot to help me *UseIntR#_2*: Based on what I just experienced - I plan to continue have this robot to help me *UseIntR#_3*: Based on what I just experienced - I would like to have this robot to help me *UseIntR#_4*: Based on what I have just experienced – I am open to have this robot to help me *UseIntR#_5R*: Based on what I have just experienced – I will not consider having this robot to help me
Trust	Engagement in the exercise (*TrustP*)	Program (robot + exercise format)	*TrustP#_1*: The following is a rating of trust in the training program. Overall, I think it's reliable *TrustP#_2*: The following is a rating of trust in the training program. On the whole, I can count on it *TrustP#_3*: The following is a rating of trust in this training program. Overall, it is capable *TrustP#_4*: The following is a rating of trust in the training program. Overall, I - am confident that it can helps me *TrustP#_5R*: The following is a rating of trust in the training program. Overall, it is unreliable (R)
TrustBase			*TrustBase#: What is your trust based on? 1. Robot's performance; 2. The interaction experience*

Satisfaction	Usefulness of the program (*UseP*)	Program (robot + exercise format)	*UseP#_1*: I think such a training program is useful *UseP#_2R*: I think such training program is not valuable(R) *UseP#_3*: I think such training programs is good for me *UseP#_4R*: I think this training program change nothing (R) *UseP#_5*: I think this training program is beneficial
Manipulation check			*Check_EngImg#:* The extent to which I actively imagined the scene described by [name of robot] in the process was *Check_EaseImg#:* The degree to which I have no trouble imagining these scenarios is: *Check_EngExcer#:* In general, the degree to which I actively participate in this training is as follows: *Check_SatR#:* In general, my satisfaction with [name of robot]'s performance is as follows: *Check_TolR#:* In general, my tolerance to robot's fault is as follows: *Check_ErrFrequent#*: In general, the frequency of robot error in the whole experiment is:

Note. # in variable name represent session

7.2.4.3 Reliability and Validity Analyses

Before conducting reliability analyses, the counter-indicative items across three timepoints (**Time**) were reversed-coded for two Affordance items (*AffEase#_4R* and *AffEase#_6R*), one Relevance item (*RelEase#_5R*), two Valence items (*ValEase#_3R* and *ValeEase#_4R*), one Use Intention item (*UseIntR#_5R*), one Trust item (*TrustP#_5R*) and two Usefulness items (*UseP#_2R* and *UseP#_4R*).

Subsequently, Cronbach's α were calculated on scales with the items within their respective scales to test scale reliability, followed by Principal Component Analysis (PCA) to test the construct validity.

7.2.4.4 Reliability and Validity Analyses with T1 data

7.2.4.4.1 Reliability Analysis

The reliability analysis (Table 7.2) showed that the *T1* experiential variable scales had high internal consistency. Problematic items (*AffEase1_4R*, *ValEase1_4R*, *UseP1_2R*, and *UseP1_4R*) were excluded from scales based on low item-total correlations until the

remaining items achieved an acceptable level of reliability. As seen in Table 7.1, Cronbach's α coefficients ranged from .790 to .940 across the scales. All included scale items loaded together effectively, indicating they assessed the same underlying constructs. Affordance of ease-of-use interaction (*AffEase1*) comprised 5 items and had an α of .790. Relevance ease-of-use (*RelEase1*; 5 items) had an alpha of .844. Valence of ease-of-use interaction (*ValEase1*) included 4 items and demonstrated an α of .915. Robots use intention (*UseIntR1*; 5 items) had the highest reliability with an α of .940. Trust in the program (*TrustP1*) included 5 items and had an α of .913. Finally, perceived usefulness (*UseP1*) contained 3 items and had an α of .927.

Table 7. 2 Reliability Analysis of Experiential Variables for *T1* in Study 3

Scale	Num of items	Items	Alpha / r	Standardized Alpha	Item mean	Variances of Item mean	*N*
AffEase1	5	*AffEase1_1, AffEase1_2, AffEase1_3, AffEase1_5, AffEase1_6R,*	.790	.795	4.555	.821	44
RelEase1	5	*RelEase1_1, RelEase1_2, RelEase1_3, RelEase1_4, RelEase1_5R*	.844	.844	4.123	1.025	44
ValEase1	4	*ValEase1_1, ValEase1_2, ValEase1_3R, ValEase1_5*	.915	.918	4.182	1.057	44
UseIntR1	5	*UseIntR1_1, UseIntR1_2, UseIntR1_3, UseIntR1_4, UseIntR1_5R*	.940	.942	4.200	1.215	44
TrustP1	5	*TrustP1_1, TrustP1_2, TrustP1_3, TrustP1_4R, TrustP1_5R*	.913	.913	4.118	.905	44
UseP1	3	*UseP1_1*, *UseP1_3*, *UseP1_5*	.927	.928	4.023	1.169	44

Note. Counter-indicative items were reverse-coded prior to analysis. Problematic items were removed from scales based on low item-total correlations. *AffEase1_4R*, *ValEase1_4R*, *UseP1_2R*, and *UseP1_4R* were excluded. $N = 44$.

7.2.4.4.2 Validity Analysis

The experiential variable items in *T1* were subjected to a principal component analysis (PCA) to examine the factor structure. Initial results based on all items showed a

Kaiser-Meyer-Olkin (KMO) measure of .785 and a significant Bartlett's test, $\chi^2(253) = 1181.01$, $p < .001$, indicating the adequacy of the data for factor analysis. However, several relevance of ease-of-use interaction (*RelEase1*) items demonstrated cross-loadings (See Table 7.3), suggesting potential issues with the scale's discriminant validity.

Table 7. 3 Pattern Matrix with 5 Components on Experiential Variables for *T1* in Study 3 (First Run)

Items	Component				
	1	2	3	4	5
AffEase1_1				.527	
AffEase1_2				.501	
AffEase1_3				.818	
AffEase1_5				.862	
AffEase1_6R				.731	
RelEase1_1					.705
RelEase1_2					.668
RelEase1_3	.513				.475
RelEase1_4			.503		
RelEase1_5R	.745				
ValEase1_1	.615				
ValEase1_2	.931				
ValEase1_3R	.830				
ValEase1_5	.893				
UseIntR1_1		.876			
UseIntR1_2		.967			
UseIntR1_3		.880			
UseIntR1_4		.754			
UseIntR1_5R		.764			
TrustP1_1			.740		
TrustP1_2			.951		
TrustP1_3			.778		
TrustP1_4R			.550		
TrustP1_5R			.755		
UseP1_1	.724				
UseP1_2	.593				
UseP1_5	.690				

After removing problematic *RelEase1* items, the PCA was re-run (Table 7.4). The second PCA with remaining items yielded a KMO of .810 and significant Bartlett's test,

$\chi^2(171) = 1003.40$, $p < .001$, confirming the appropriateness of the factor analysis. A clear five-component structure emerged, accounting for 78.21% of the total variance (Table 7.4)

The 4 *ValEase1* items loaded strongly on the first component (range .69 to .97), while the 5 *UseIntR1* items showed strong loadings on the second component (range .77 to 1.01). The third component comprised the 5 *TrustP1* items with loadings ranging from .56 to .96. The 5 *AffEase1* items demonstrated strong loadings on the fourth component (range .50 to .86), and the two remaining *RelEase1* items loaded onto the fifth component (loadings of .60 and .74). Interestingly, the 3 *UseP1* items loaded together on the first component (range .52 to .63), indicating a high correlation with the *ValEase1* scale.

Table 7. 4 Pattern Matrix with 8 Components on Experiential Variables for *T1* in Study 3 (Second Run)

Items	Component				
	1	2	3	4	5
AffEase1_1				.527	
AffEase1_2				.501	
AffEase1_3				.818	
AffEase1_5				.862	
AffEase1_6R				.731	
RelEase1_1					.743
RelEase1_2	**.512**				.598
ValEase1_1	.693				
ValEase1_2	.971				
ValEase1_3R	.781				
ValEase1_5	.972				
UseIntR1_1		.901			
UseIntR1_2		1.008			
UseIntR1_3		.936			
UseIntR1_4		.800			
UseIntR1_5R		.771			
TrustP1_1			.747		
TrustP1_2			.956		
TrustP1_3			.803		
TrustP1_4R			.559		
TrustP1_5R			.782		
UseP1_1	.608				
UseP1_2	.519				
UseP1_5	.629				

7.2.4.4.3 Reliability and Validity Analyses with T2 and T3 data

Principal component analyses (PCA) were conducted on *T2* and *T3* experiential variable items that achieve high item-total correlations within their own scale. The PCA aimed to validate the factor structure across timepoints.

For *T2* items remaining after reliability analyses that achieved high internal consistency within their scales, the initial PCA yielded a KMO of .812 and significant Bartlett's test, $\chi^2(210) = 1144.20$, $p < .001$. However, some *UseP2* and *RelEase2* items showed multiple loadings (See Table 7.5), so I conducted another PCA after removing the *UseP2* and *RelEase2* items corresponding to the removed items in *T1*. The second PCA on remaining *T2* items resulted in a KMO of .829 and Bartlett's $\chi^2(136) = 821.85$, $p < .001$, explaining 79.78% of variance. As expected, and shown in Table 7.6, items loaded onto distinct Use Intention to Robot and Usefulness of the Program (component 1), Valence (component 2), Trust on the Program (component 3), and Relevance (component 4) components.

Table 7. 5 Pattern Matrix with 4 Components on Experiential Variables for T2 in Study 3 (First Run)

Items	Component				
	1	2	3	4	5
RelEase2_1				.793	
RelEase2_2				.621	
RelEase2_3					.672
RelEase2_4				.520	
RelEase2_5R	.768				
ValEase2_1		1.014			
ValEase2_2		.652			
ValEase2_3R		.978			
ValEase4_2		.750			
ValEase2_5	.466	.582			
UseIntR2_1	1.037				
UseIntR2_2	.868				
UseIntR2_3	.877				
UseIntR2_4	.580				

UseIntR2_5R	.878				
TrustP2_1		.482	.697		
TrustP2_2			.589		
TrustP2_3			.698		
TrustP2_4R			.947		
TrustP2_5R			.928		
UseP2_1	.635				
UseP2_2R	.627				-.493
UseP2_3					.535
UseP2_4R				-.614	
UseP2_5	.764				

Table 7. 6 Pattern Matrix with 4 Components on Experiential Variables for T2 in Study 3 (Second Run)

Items	Component			
	1	2	3	4
RelEase2_1				.762
RelEase2_2				.951
ValEase2_1			.922	
ValEase2_2			.582	
ValEase2_3R			1.022	
ValEase2_5			.561	
UseIntR2_1	1.016			
UseIntR2_2	.994			
UseIntR2_3	.926			
UseIntR2_4	.534			
UseIntR2_5R	.890			
TrustP2_1		.724	.535	
TrustP2_2		.647		
TrustP2_3		.730		
TrustP2_4R		.873		
TrustP2_5R		.933		
UseP2_1	.700			
UseP2_3	.686			
UseP2_5	.787			

At *T3*, the initial PCA with the items remaining after reliability analyses that achieved high internal consistency within their scales produced a KMO of .853 and significant Bartlett's test, $\chi^2(210) = 1135.85$, $p < .001$. The factor loadings are shown in Table 7.7. After removing cross-loading *RelEase3* items, the second PCA (Table 7.8) demonstrated improved a factor structure (KMO = .854; Bartlett's $\chi^2(120) = 950.26$, $p < .001$), accounting for 82.24%

of variance. However, the expected component structure did not clearly emerge. Specifically, *RealEase3_2* loaded almost equally on component 1(.551) and 4 (.574). Additionally, two *UseIntR3* items, *UseIntR3_1* and *UseIntR3_4*, diverged from other scale items on component 2 (loading of .49 and .46). *TrustP3_1* also showed near equal loadings on component 1 (.529) and 3 (.504).

Although the T*3* item structure did not remain fully consistent with the *T1* and *T2* analyses, the items were retained for group comparisons. However, interpretation regarding distinct construct measurement should be made cautiously given the limitations identified.

Table 7. 7 Pattern Matrix with 4 Components on Experiential Variables for *T3* in Study 3 (First Run)

Items	Component			
	1	2	3	4
RelEase3_1		.623		
RelEase3_2	.905			
RelEase3_3	.664			
RelEase3_4				1.000
RelEase3_5R				.738
ValEase3_1	.700			
ValEase3_2				
ValEase3_3R				
ValEase3_5	.835			
UseIntR3_1		.888		
UseIntR3_2		.955		
UseIntR3_3		.901		
UseIntR3_4		.810		
UseIntR3_5R		.595		
TrustP3_1			.521	
TrustP3_2			.720	
TrustP3_3			.869	
TrustP3_4R			.519	
TrustP3_5R			1.058	
UseP3_1				.453
UseP3_3	.521			
UseP3_ 5				.855

Table 7. 8 Pattern Matrix with 4 Components on Experiential Variables for T3 in Study 3 (Second Run)

Items	Component			
	1	2	3	4
RelEase3_1				.896
RelEase3_2	.551			.574
ValEase3_1	.969			
ValEase3_2	.868			
ValEase3_3R	.649			
ValEase3_5	.956			
UseIntR3_1	.492			
UseIntR3_2		.678		
UseIntR3_3		.776		
UseIntR3_4	.461			
UseIntR3_5R		.984		
TrustP3_1	.529		.504	
TrustP3_2			.699	
TrustP3_3			.559	
TrustP3_4R			.462	
TrustP3_5R		.493	.781	
UseP3_1	.772			
UseP3_3	.721			
UseP3_5	.493			

Reliability analyses examined the internal consistency of the experiential variable scales at *T2* and *T3* using the reduced item sets. As seen in Table 7.9, all scales showed high reliability at both time points.

At *T2*, *RelEase2* comprised 2 items (*RelEase2_1* and *RelEase2_2*) with a Spearman rho of .83. *ValEase2* included 4 items (*ValEase2_1*, *ValEase2_2*, *ValEase2_3R* and *ValEase2_5*) and yielded an α of .89. *UseIntR2* with 5 items (*UseIntR2_1*, *UseIntR2_2*, *UseIntR2_3*, *UseIntR2_4* and *UseIntR2_5R*) demonstrated an α of .94. *TrustP2* contained 5 items (*TrustP2_1*, *TrustP2_2*, *TrustP2_3*, *TrustP2_4R* and *TrustP2_5R*) and resulted in an α of .89. Finally, *UseP2* with 3 items had an α of .84.

Similarly, scales in *T3* had the same items as that in *T2* and maintained high levels of internal consistency. *RelEase3* had a Spearman rho of .83 with 2 items; *ValEase3* had an α of .94 with 4 items; *UseIntR3* produced an α of .95 with 5 items; *TrustP3* yielded an α of .91 with 5 items; and *UseP3* demonstrated an α of .92 with 3 items.

Table 7. 9 Reliability Analyses of Scales for T2 and T3 in Study 3

Scale	Num of items	Items	Alpha / r	Standardized Alpha	Item mean	Variances of Item mean	*N*
			Time = 2				
RelEase2	5	*RelEase2_1, RelEase2_2*	r = .828, p <.001				44
ValEase2	4	*ValEase2_1, ValEase2_2, ValEase2_3R, ValEase2_5*	.891	.895	4.239	1.030	44
UseIntR2	5	*UseIntR2_1, UseIntR2_2, UseIntR2_3, UseIntR2_4, UseIntR2_5R*	.942	.984	4.291	1.366	44
TrustP2	5	*TrustP2_1, TrustP2_2, TrustP2_3, TrustP2_4R, TrustP2_5R*	.893	.898	4.218	1.229	44
UseP2	3	*UseP2_1*, *UseP2_3*, *UseP2_5*	.836	.839	4.242	1.258	44
			Time = 3				
RelEase3	5	*RelEase3_1, RelEase3_2*	r = .720, p <.001				44
ValEase3	4	*ValEase3_1, ValEase3_2, ValEase3_3R, ValEase3_5*	.943	.943	4.278	1.292	44
UseIntR3	5	*UseIntR3_1, UseIntR3_2, UseIntR3_3 UseIntR3_4, UseIntR3_5R*	.946	.948	4.445	1.036	44
TrustP3	5	*TrustP3_1, TrustP3_2, TrustP3_3, TrustP3_4R, TrustP3_5R*	.912	.915	4.464	.926	44
UseP3	3	*UseP3_1*, *UseP3_3*, *UseP3_5*	.924	.924	4.371	1.191	44

7.2.4.5 Outliers Exploration

7.2.4.5.1 Exploration of Experiential Outliers

Table 7.10 and Table 7.11 present the distribution of outliers across the three sessions for both the *Audio* and *Avatar* conditions. To identify potential outliers, means were calculated across items for the six experiential scales at each of the three timepoints. The resulting 16 mean values were prefixed with "*M_*" to distinguish them from single-item values. Outlier analyses were then performed on each scale using boxplots with **Medium** factors.

In the initial analysis ($N = 44$), several participants were identified as potential outliers across various measures and sessions (Table 7.10). Outliers in the lower end of the box are marked with brackets, and extreme outliers are denoted with an asterisk. Participants E3_19, E3_20, E3_36, and E3_41were excluded due to specific reasons, such as inability to speak, harsh responses, slow performance, or expressing discomfort with the experiment. Consequently, the sample size was reduced to 40 participants ($n = 40$) for further exploration.

Table 7. 10 Outlier Distribution for Experiential Variables Across Medium and Time for N = 44 in Study 3

Condition	**Audio**			**Avatar**		
Session	**1**	**2**	**3**	**1**	**2**	**3**
M_AffEase	E3_5, E3_14, (E3_19*)			(E3_36)		
M_RelEase	E3_13, (E3_19)	/	/	(E3_36)	/	/
M_ValEase	(E3_19, E3_20*)	/	(E3_19, E3_20)	(E3_36, E3_43)	/	/
M_UseIntP	E3_10, E3_13, (E3_19*, E3_20)	(E3_19)	(E3_19)	/	(E3_25)	(E3_25)
M_TrustP	(E3_19)	(E3_19)	(E3_19)	(E3_41)	E3_31, E3_40, (E3_36, E3_41*)	E3_31, E3_37, E3_40, (E3_41)
M_UseP	(E3_19*, E3_20*)	(E3_19)	(E3_20)	(E3_41, E3_25)	/	(E3_41)

Note. Outliers in the lower end of the box are marked with a bracket.

A subsequent analysis of outliers with the remaining 40 ($n = 40$) participants revealed a smaller number of outliers (Table 7.11). The outliers were primarily concentrated in the lower end of the box, with a few extreme outliers marked with an asterisk. As the Generalized Estimating Equations (GEE) approach considers outliers in its analysis, these participants were not excluded from the final sample.

Table 7. 11 Outlier Distribution for Experiential Variables Across Medium and Time for n = 40 in Study 3

Condition	**Audio**			**Avatar**		
Session	**1**	**2**	**3**	**1**	**2**	**3**
M_AffEase	E3_5, E3_14, (E3_4)			/		
M_RelEase	/	/	(E3_18)	/	/	/

M_ValEase	/	/	E3_10	E3_43	/	/
M_UseIntP	E3_10*, E3_13	/	/	/	(E3_25)	(E3_25)
M_TrustP	/	(E3_3)	/	/	E3_31	E3_31, E3_37, E3_40, (E3_33)
M_UseP	/	/	/	E3_25	/	/

Note. Outliers in the lower end of the box are marked with a bracket.

7.2.4.5.2 Exploration of Assessment Outcome Outliers

To ensure the accuracy of the intervention outcome analyses, outliers were identified in each assessment by categorizing participants with non-serious conditions and low effects (Chapter 4). The *Control* group included participants from Chapter 4. Table 7.12 indicates the distribution of outliers across assessments.

For BII-D2, three participants exhibited an unexpected substantial decrease in depressive scores. Participant E3_7's BDI-II score dropped remarkably from 35 to 1. As she reported benefiting greatly from the exercise by changing her thinking mode and feeling relieved, she was retained in the analysis pool. Participant E3_22's BDI-II score decreased from 31 to 0, but he changed exercise schedules frequently due to illness, raising concerns about response validity. Participant E3_40's BDI-II score reduced substantially from 23 to 0 despite good collaboration; she attributed this decrease to recovery from illness rather than the exercise. These two participants, along with participant C_18 from the *Control* group who exhibited an extreme BDI-II score reduction, were regarded as BDI-II outliers requiring cautious analysis due to possible confounding factors.

For SST, two non-serious participants were identified for pre-SST (E3_2, E3_36) and one for post SST (E3_25). C_13 from the *Control* groups were also non-serious in SST.

When comes to WSAP, polarized response biases in pre WSAP were found in 8 participants (E3_1, E3_2, E3_6, E3_7, E3_19, E3_26, E3_36 and E3_37) and in post WSAP for 8 participants (E3_2, E3_12, E3_18, E3_25, E3_32, E3_37, E3_39 and E3_41). E3_2 and E3_37 showed consistent disagreement on two sets of materials (sentences with positive or

negative wording) in pre and post WSAP. E3_41 consistently disagreed with the first set of materials but consistently agreed with the second set.

In pre WSAP, E3_6 consistently disagreed with both sets of materials, while E3_1, E3_26, and E3_36 consistently disagreed with the first set. E3_7 and E3_19 consistently agreed with the first set. In post WSAP, E3_12, E3_18, E3_25, E3_31, and E3_39 consistently disagreed with the first set.

Regarding non-serious participants, E3_36 less agreed with positive stimuli in Set 1 but more agreed in Set 2 (6-point difference) in pre-WSAP, whereas E3_41 less agreed with negative stimuli in Set 1 but more agreed with in Set 2 (7 score difference). In post-WSAP, E3_12 and E3_41 less agreed with positive stimuli in Set 1 but more in Set 2 (6- and 7-point differences, respectively), while E3_18 and E3_26 less agreed with negative stimuli in Set 1 but more in Set 2 (7- and 6-point differences).

Taken together, participants E3_1, E3_6, E3_7, and E3_19, who consistently disagreed or agreed with Set 1 and/or Set 2 in pre WSAP, and E3_25, E3_32, and E3_39, who consistently disagreed with Set 1 in post WSAP, were not considered outliers as they might have been overly cautious about endorsing the correlation between sentences and wording stimuli. The remaining participants were regarded as WSAP outliers due to differences in endorsement or rejection of specific polarized stimuli in both sets or consistent rejection or endorsement across sets and timepoints. C_11 and C_4 from Study 1 remained as WSAP outliers.

In pre SRT, six participants (E3_4, E3_7, E3_12, E3_19, E3_29, and E3_37) were identified as outliers based on response patterns with bias 2 and 4. In post-SRT, two participants (E3_42 and E3_43) were outliers based on the same criteria. Additionally, E3_2 and E3_38 were considered outliers for answering more than 5 comprehension questions incorrectly. C_4 and C_16 from Chapter 4 remained as SRT outliers.

Table 7.13 presents the distribution of outliers across the **Medium** conditions. The sample sizes are reported both with and without outliers for each assessment and condition.

Table 7. 12 Outlier Distribution for SST, WSAP, and SRT in Study 3

Participant index	Pre-test				Post-test			
	Non-serious participants in SST [1]	Non-serious participants in WSAP [2]	Non-serious participants in SRT [3]	Low effect participants in SRT [4]	Non-serious participants in SST	Non-serious participants in WSAP	Non-serious participants in SRT	Low effect participants in SRT
E3_1	/	Set 1, disagree	/	/	/	/	/	bias 1
E3_2	6 incorrect	Set 12, disagree	6 incorrect	/	/	Set 12, disagree	7 incorrect	/
E3_3	/	/	/	/	/	/	/	bias 1
E3_4	/	/	6 incorrect	bias 4	/	/	/	/
E3_5	/	/	/	/	/	/	/	bias 1
E3_6	/	Set 12, disagree	/	/	/	/	/	/
E3_7	/	Set 1, agree,	7 incorrect	bias 2	/	/	/	/
E3_8	/	/	/	/	/	/	/	bias 1
E3_12	/	/	/	bias2	/	Set 1, disagree, PE_D: 6	/	/
E3_13	/	/	/	/	/	/	/	bias 1
E3_14	/	/	/	/	/	/	/	bias 1
E3_15	/	/	/	/	/	/	/	bias 1
E3_18	/	/	/	/	/	Set 1, disagree, NM_D: -7	/	/
E3_19	/	Set 1, agree	7 incorrect	bias 2	/	/	6 incorrect	/
E3_22	/	/	/	/	/	/	/	bias 1
E3_24	/	/	/	/	/	/	/	bias 1
E3_25	/	/	/	bias 1	7 incorrect	Set 1, disagree	/	/
E3_26	/	Set 1, disagree	/	/	/	NM_D: -6	/	/
E3_27	/	/	/	/	/	/	/	bias 1
E3_28	/	/	/	/	/	/	/	bias 1
E3_29	/	/	/	bias 2	/	/	6 incorrect	bias 1
E3_30	/	/	/	/	/	/	/	bias 1
E3_31	/	/	/	/	/	/	/	bias 1
E3_32	/	/	/	/	/	Set 1, disagree	/	/
E3_34	/	/						
E3_36	7 incorrect	Set 1, disagree. PE_D: 6	/	/	/	/	/	/
E3_37	/	Set 12, disagree	7 incorrect	bias 2	/	Set 12, disagree	/	/
E3_38	/	/	/	/	/	/	7 incorrect	/
E3_39	/	/	/	/	/	Set 1, disagree	/	/
E3_40	/	/	/	bias 1	/	/	/	bias 1
E3_41	/	NM_D: -7	/	/	/	Set 1, disagree, Set 2, agree, PE_D: 7	/	/
E3_42	/	/	/	/	/	/	/	bias 2
E3_43	/	/	7 incorrect	/	/	/	/	bias 2
E3_44	/	/	/	bias 1	/	/	/	bias 1
C_4	/	/	/	/	/	PE_D: 10	/	bias 2
C_7	/	/	/	/	/	/	/	
C_11	/	PE_D: 8	/	/	/	/	/	/
C_12	/			bias 1				
C_13	/	/	6 incorrect	/	/	/	/	/
C_14	/	/	/	bias 1	/	/	/	/
C_16	/	/	6 incorrect	bias 2	/	/	4 incorrect 3 null	/

Note. NM_D means the rejection difference on negative words in two sets materials; PE_D means the endorsement difference on positive words in two sets materials; bias 1 means acceptance on positive statements without considering the context; bias 2 means unbiased rejection without considering the context; bias 3 means unbiased acceptance without considering context; bias 4 means acceptance on negative statements without considering context.

Table 7. 13 Outlier Distribution Across Medium for SST, WSAP, and SRT in Study 3

Items	Medium	Sample Size (total)	Outliers Index (Pre)	Outliers Index (post)	Sample Size (without outliers)
DS_MS	Audio	23	/	E3_22	22
	Avatar	21	/	E3_40	20
	Control	18	/	C_18	17
SST	Audio	23	E3_2	/	22
	Avatar	21	E3_36	E3_25	19
	Control	18	C_13	/	17
SRT	Audio	23	E3_2, E3_4, E3_7, E3_12, E3_19	E3_2, E3_19	18
	Avatar	21	E3_29, E3_37, E3_43	E3_29, E3_38, E3_42, E3_43	16
	Control	18	C_16	C_4	16
WSAP	Audio	23	E3_2	E3_2, E3_12, E3_18	20
	Avatar	21	E3_36, E3_37, E3_41	E3_26, E3_37, E3_41	17
	Control	18	C_11	C_4	16

7.3 Results

7.3.1 Demographics

Continuous *Age* variable and *Depression* variable were recoded into categorical variables *Age range* and *Depression Level*, respectively, for Chi-square test. The demographic distribution was depicted in Table 7.14. I observed the Table 7.14 that over 20% of cells had cells had expected counts less than five, violating assumption of Pearson Chi-square.

Table 7. 14 Demographic Distribution Over Medium for N = 40 in Study 3

Demographics	Audio (n = 21)	Avatar (n = 19)	Overall (n = 40)
Gender			
Female	16 (16.8)	16 (15.2)	32
Male	5 (4.2)	3 (3.8)	8

Depression Level			
0-14: Minimal	2 (1.6)	1 (1.4)	3
15-20: Mild	0 (0.5)	1 (0.5)	1
21-29: Moderate	8 (9.5)	10 (8.6)	18
30-63: Severe	11 (9.5)	7 (8.6)	18
Age (years)			
20 and below	4 (3.2)	2 (2.9)	6
21-24	7 (7.4)	7 (6.7)	14
25-29	8 (7.9)	7 (7.1)	15
30 and above	2 (2.6)	3 (2.4)	5

Note. Expected frequencies for each cell.

Monte Carlto simulations, instead, assessed random assignment adequacy across *Gender*, *Age range*, and *Depression level*. Results (Table 7.15) indicated no significant associations for *Depression Level* by **Medium**, $p = .550$ (95% CI [.537, .563]); and *Age Range* by **Medium**, $p = .884$ (95% CI [.876, .892]). The expected result was provided for *Gender* by **Medium** was .698. These suggested successful random assignments across examined demographics.

Table 7. 15 Chi-Square Value on Age Range, Gender, and Depression Level Across Medium in Study 3

	Value	df	Asymptotic Significance (2-sided)	Pearson Chi-square Assumption [a]	Monte Carlo Sig. (2-sided)		
					Significance	95% Confidence Interval	
						Lower Bound	Upper Bound
Gender × Medium	.401	1	.527	Violated (50.0%)	.698[b]	/	/
Depression Level × Medium	2.350	3	.503	Violated (50.0%)	.550	.537	.563
Age Range × Medium	.835	3	.841	Violated (50.0%)	.884	.876	.892

a. The assumption concerns the percentage of cells that have an expected count of less than five.
b. For 2×2 crosstabulation, exact results are provided instead of Monte Carlo results.

7.3.2 Manipulation Check

Table 7.16 displays the manipulation check ranks distribution across three timepoints (*T1- T3*) for *Audio* and *Avatar* conditions. The ranks were categorized as *Low* (1-4), *Medium* (5-7), and *High* (8-10). Six variables were calculated: ease of imaging scenarios (*EaseImg_Rank*), active engagement in imagining scenarios (*EngImg_Rank*), active participation in training (*EngExcer_Rank*), satisfaction with robot performance (*SatR_Rank*),

tolerance of robot errors (*TolR_Rank*), and perceived frequency of robot errors (*ErrFrequent_Rank*). The number of participants in each rank category is reported.

Table 7. 16 Manipulation Checks Over Time (T1-3) Measured as Rank Data (1-4 as Low, 5-7 as Medium, and 8-10 as High), Number of Participants in Each Rank, $N = 40$ in Study 3

Measures		*T1*			*T2*			*T3*		
		Low	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
1. *EngImg_Rank*	Audio	0	4	17	1	4	16	0	2	19
	Avatar	2	3	14	1	3	15	0	5	14
2. *EaseImg_Rank*	Audio	2	5	14	2	5	14	0	6	15
	Avatar	2	2	15	1	6	12	0	1	18
3. *EngExcer_Rank*	Audio	1	3	17	2	6	13	2	6	13
	Avatar	0	6	13	1	5	13	0	6	13
4. *SatR_Rank*	Audio	0	7	14	1	5	15	0	6	15
	Avatar	1	4	14	0	6	13	1	5	13
5. *TolR_Rank*	Audio	0	7	14	1	11	9	/	/	/
	Avatar	2	2	15	0	3	16	/	/	/
6. *ErrFrequent_Rank*	Audio	/	/	/	/	/	/	9	7	5
	Avatar	/	/	/	/	/	/	11	5	3

Note: 1. EngImg_Rank: The rank of the participant actively imagined the scenario.
2. EaseImg_Rank: The rank of the ease of imagining these scenarios.
3. EngExcer_Rank: The rank of the participants actively participates in this training.
4. SatR_Rank: The rank of participant's satisfaction with robot's performance.
5. TolR: The rank of participant's tolerance to robot's fault.
6. ErrFrequent_Rank: The rank of participant thought the error the robot made during six sessions.

Table 7.17 presents trends in four rank variables (*EngImg_Rank*, *EaseImg_Rank*, *EngExcer_Rank*, *SatR_Rank*) across *T1- T3*. The trends summarize changes in ranks over time as staying *High*, dropping from *High* to *Medium*, etc. The number of participants exhibiting each trend is provided, along with participant codes in brackets. The *Audio* and *Avatar* conditions are compared. Variables showed predominantly *High-High-High* trends, indicating consistent *High* ranks across timepoints. Other trends represent variations in ranks across time.

Table 7. 17 Rank Trend Over Time (T1-3), Number of Participants with Each Trend, Index in Bracket Represents Participants with That Trend, n = 40 in Study 3

Trend across *T1*, *T2* and *T3*	1. *EngImg_Rank*		2. *EaseImg_Rank*		3. *EngExcer_Rank*		4. *SatR_Rank*	
	Audio	Avatar	Audio	Avatar	Audio	Avatar	Audio	Avatar
High-High-High	14	13	11	9	11	11	12	11
High-High-Medium	0	1 (E3_30)	/	/	2 (E3_14, E3_21)	0	1 (E3_13)	1 (E3_30)

High-Medium-High	2 (E3_1, E3_2)	0	2 (E3_3, E3_13)	5 (E3_31, E3_39, E3_42, E3_43, E3_28)	1 (E3_15)	1 (E3_39)	1 (E3_15)	1 (E3_34)
High-Medium-Medium	1 (E3_15)	0	1 (E3_14)	0	3 (E3_2, E3_5, E3_9)	1 (E3_28)	1 (E3_14)	0
High-Medium-Low	/	/	/	/	1 (E3_13)	0	0	1 (E3_33)
High-Low-Medium	0	1 (E3_33)	1 (E3_5)	1 (E3_30)	0	1 (E3_33)	/	/
Medium-High-High	2 (E3_21, E3_23)	1 (E3_37)	2 (E3_6, E3_11)	2 (E3_26, E3_29)	1 (E3_23)	2 (E3_38, E3_42)	2 (E3_7, E3_23)	0
Medium-High-Medium	/	/	1 (E3_23)	0	0	1 (E3_44)	1 (E3_16)	1 (E3_28)
Medium-Medium-High	1 (E3_4)	0	/	/	1 (E3_22)	0	1 (E3_4)	1 (E3_44)
Medium-Medium-Medium	0	2 (E3_43, E3_44)	2 (E3_2, E3_4)	0	0	3 (E3_25, E3_30, E3_43)	2 (E3_2, E3_5)	2 (E3_25, E3_43)
Medium-Low-Medium	1 (E3_5)	0	/	/	1 (E3_18)	0	1 (E3_3)	0
Low-High-High	0	1 (E3_28)	1 (E3_7)	1 (E3_37)	/	/	/	/
Low-Medium-High	/	/	0	1 (E3_34)	/	/	/	/
Low-Medium-Medium	0	1 (E3_25)	/	/	/	/	0	1 (E3_42)
Low-Low- Medium	/	/	1 (E3_18)	0	/	/	/	/
Low-Low-Low	/	0	/	/	2 (E3_16)	0	/	/
Total	21	19	21	19	21	19	21	19

Note. 1.EngImg_Rank: The rank of the participant actively imagined the scenario; 2. EaseImg_Rank: The rank of the ease of imagining these scenarios; 3. EngExcer_Rank: The rank of the participants actively participates in this training; 4. SatR _Rank: The rank of participant's satisfaction with robot's performance

In Table 7.16, most participants reported relatively high engagement in imagery (*EngImg_Rank*) across timepoints, though active participation in the exercise (*EngExcer_Rank*) was more moderate overall, indicating the manipulation on **Medium** type did not significantly affect the active engagement and ease of imagery. Several participants initially reported difficulty with ease of imagining scenarios (*EaseImg_Rank*), but by *T3* nearly all participants in the *Avatar* condition reported no issues. Regarding robot satisfaction (*SatR_Rank*), approximately three-fourths conveyed satisfaction. Tolerance of robot errors (*TolR_Rank*) appeared higher in the avatar condition, with 16 ranking it as *High*. On

perceived error frequency (*ErrFrequent_Rank*), over half of the avatar condition ranked it as *Low*, compared to under half for *Audio*.

Examining trends in Table 7.17, over half maintained high *EngImg_Rank* and *EngExcer_Rank*. Similarly, over half displayed a consistent high *SatR_Rank* trend. Approximately half showed a high *EaseImg_Rank* trend across timepoints. Low or dropping trends in these variables, including *High-Medium-Low* (↓)*, High-Low-Medium* (↓)*, Medium-Medium-Medium* (→), *Medium-Low-Medium* (→), *Low-Medium-Medium* (→), *Low-Low-Medium* (→), *Low-Low-Low* (→), occurred at similar rates (last column in Table 7.17) in both conditions.

Given the influence of *EngImg_Rank* and *EngExcer_Rank* on outcomes (Hoppitt et al., 2010; Mathews & Mackintosh, 2000), participants showing decreasing trends and constant low trends for these variables can be considered engagement outliers (E3_5, E3_13, E3_16, E3_18, E3_25, E3_30, E3_33, E3_43, E3_44). Those with low *EaseImg_Rank* trends (E3_2, E3_4, E3_18) are ease-imagery outliers. Low and decreasing satisfaction trends indicate satisfaction outliers (E3_2, E3_3, E3_5, E3_25, E3_33, E3_42, E3_43) [3]. Further analyses were performed with datasets of n = 40, *n (without engagement outliers)* = 31, *n (without ease-imagery outliers)* = 37, and *n (without satisfaction outliers)* = 33 to inspect differences across time among conditions.

Pearson correlation between participant's tolerance to robot mistake (*Check_Tol1*, *Check_Tol2*) and the perceived robot mistakes (*Check_ErrFrequent*) were performed using the frequency data. Results showed no significant correlations between tolerance and perceived robot mistakes (*Check_Tol1:* $r = .188$, $p = .245$; *Check_Tol2:* $r = -.033$, $p = .838$) or between the change in *Check_TolR* (*Check_Tol2- Check_Tol1*) and *Check_ErrFrequent* ($r = -.161$, $p = .321$). This suggests that participants' tolerance and tolerance changes were not necessarily related to how often they felt the robot was making errors. Individual differences

in the perception of errors and disposition towards errors might influence their perception of mistakes made by the robots (Salem et al., 2015; Mirnig et al., 2017), hindering a direct correlation between these indicators. The lack of change in tolerance based on perceived mistakes might be due to real-time justification of errors during the exercise, as inspired by Klüber and Onnasch (2022), who found that providing reasonable justifications and acknowledging errors can maintain participants' trust and mitigate negative effects. The **eiIBM_RobotV2** robots quickly apologize and correct its wordings when making mistakes, which could avoid the user's negative emotion and reduced tolerance.

A two-way logistic regression examined whether tolerance of mistake (*Check_Tol1*, *Check_Tol2*) moderated the effect of perceived mistake frequency (*Check_ErrFrequent*) on participant satisfaction (coded as 0 for high and/or consistent satisfaction and 1 for low and/or decreasing satisfaction). The model with interaction terms was statistically significant, $\chi^2(2) = 6.974$, $p = .031$, explaining 16% to 26.8% variance based on Cox & Snell R^2 and Nagelkerke R^2.

Results showed a significant interaction between *Check_ErrFrequent* and initial tolerance (*Check_Tol1*) in predicting satisfaction trend ($B = -.153$, $p = .048$). The odds ratio indicated that for every one unit increase in *Check_ErrFrequent* × *Check_Tol1* interaction term, the odds of being in the low satisfaction trend decreased by 14.2%. There was no significant interaction with later tolerance ratings (*Check_Tol2*).

These findings provide preliminary evidence that higher initial tolerance mitigates the negative impact of perceiving more frequent robot mistakes on satisfaction. Participants with greater tolerance may maintain higher satisfaction despite noticing more mistakes, possibly because their negative emotional responses (i.e., frustration or annoyance) to the robot's errors are less compared to non-tolerant participants.

7.3.3 Descriptive Statistics of Experiential Variables

Table 18 presents descriptive statistics for eight experiential variables across two **Medium** (*Audio* and *Avatar*) at three different timepoints (*T1, T2, T3*), n = 40., with a total sample size of 40. The mean values are reported in each cell, followed by the standard deviations in parentheses. The number of observations (Num) for each medium at each time point is also provided at the bottom of the table. Cells in bold indicate the highest value across mediums, an asterisk (*) indicates a higher value than the previous timepoint(s), and a double asterisk (**) indicates a higher value than the previous two timepoints.

Several noteworthy patterns emerge from the data. First, the variable *M_AffEase* was only measured at *T1*, with identical mean values of 4.64 for both Audio and Avatar. Second, *M_RelEase* was measured at all time points, with Avatar consistently showing higher mean values than *Audio* at *T2* and *T3*. Third, *M_ValEase* reached its highest mean value of 4.77 for Audio at *T3*, surpassing the values at the previous two time points. Fourth, *M_UseIntR* showed no substantial differences between *Audio* and *Avatar* at *T3*, with both mediums exhibited an increasing trend over time. Fifth, *M_TrustP* achieved its highest mean value of 4.65 for *Avatar* at *T3*, exceeding the values for *Audio* at all time points. Finally, *M_UseP* reached its peak mean value of 4.63 for Audio at *T3*, surpassing the values for *Avatar* at all time points.

Although these descriptive statistics did not measure significant differences, they provided tendency to the researchers that audio elicited a more functional evaluation (usefulness) while the avatar elicited a more affective evaluation (trust). The avatar received the highest use intention compared to audio across all time points.

Table 7. 18 Descriptive Statistics of Experiential Variables Across Medium and Time in Study 3

Measures	*T1*		*T2*		*T3*	
	Audio	*Avatar*	*Audio*	*Avatar*	*Audio*	*Avatar*
M_AffEase	4.64 (.539)	4.64 (.465)	/	/	/	/
M_RelEase	**4.43 (.712)**	4.31 (.803)	4.33 (1.053)	**4.66 (.898) ***	4.45 (.921)	**4.63 (.926)**
M_ValEase	4.38 (.551)	**4.41 (.723)**	**4.37 (.769) ***	4.30 (.836)	**4.77 (.666) ****	4.43 (.920)
M_UseIntR	4.28 (.608)	**4.48 (1.01)**	4.41 (.733) *	**4.55 (1.015) ***	**4.58 (.712) ****	**4.58 (.975) ****
M_TrustP	4.20 (.693)	**4.34 (.608)**	4.28 (.791) *	**4.44 (.659) ***	4.53 (.676) **	**4.65 (.699)**
M_UseP	**4.33 (.494)**	4.18 (.856)	**4.49 (.712) ***	4.35 (.885) *	**4.63 (.690)**	4.44 (.975) **
Num	19	21	19	21	19	21

Note. Mean (SD) in each cell. Highest mean across Medium was in bold. An asterisk (*) indicates a higher value than the previous timepoint(s), and a double asterisk (**) indicates a higher value than the previous two timepoints.

7.3.4 Correlation Between Experiential Variables and Demographics

Pearson correlations were conducted to examine the relationships between demographics (*Gender*, *Age*), *Depression*, and experiential variables across timepoints (*T1*, *T2*, *T3*). Depression showed significant negative correlations with several experience variables (Table 7.19).

At *T1*, higher depression is related to lower program trust (*M_TrustP1*: r = -.41, p =.008). At *T2*, depression is associated with lower ease-of-use relevance (*M_RelEase2*: r = -.38, p =.017), valence (*M_ValEase2*: r = -.44, p =.004), and robot use intention (*M_UseIntR2*: r = -.41, p =.009). Higher depression also correlated with lower trust at *T2* (*M_TrustP2*; r = -.39, p =.013).

Similarly, at *T3*, increased depression correlated with reduced ease-of-use relevance (*M_RelEase3*; r = -.39, p = .014), valence (*M_ValEase3*: r = -.31, p = .05), robot use intention (*M_UseIntR3*: r = -.33, p = .041), and perceived usefulness (*M_UseP3*: r = -.33, p =.039). Therefore, *Depression* or the *Depression Level* were added as the covariates in GEE or MANOVA models.

Table 7. 19 Correlation Between Demographic Variables and Experiential Variables in Study 3

	Gender	*Age*	*Depression*

M_AffEase1	-.015 (.926)	-.003 (.838)	-.010 (.949)
M_RelEase1	-.338* (.033)	-.065 (.691)	-.070 (.669)
M_ValEase1	-.141 (.387)	.016 (.922)	-.274 (.088)
M_UseIntR1	-.015 (.924)	.270 (.092)	-.295 (.065)
M_TrustP1	.008 (.962)	-.118 (.470)	-.414** (.008)
M_UseP1	.117 (.472)	.146 (.368)	-.266 (.097)
M_RelEase2	.039 (.813)	-.051 (.756)	-.375* (.017)
M_ValEase2	-.056 (.732)	.030 (.855)	-.442** (.004)
M_UseIntR2	.044 (.789)	.152 (.349)	-.407** (.009)
M_TrustP2	-.073 (.654)	-.114 (.486)	-.388* (.013)
M_UseP2	-.165 (.308)	-.013 (.939)	-.304 (.057)
M_RelEase3	.014 (.933)	.135 (.405)	-.385* (.014)
M_ValEase3	-.072 (.657)	.095 (.559)	-.312* (.050)
M_UseIntR3	.118 (.468)	.267 (.095)	-.325* (.041)
M_TrustP3	-.086 (.600)	.037 (.823)	-.277 (.084)
M_UseP3	-.051 (.756)	.226 (.161)	-.327* (.039)

7.3.5 Generalized Estimating Equations on Experiential Variables

Although most of the data exhibited normal distributions (Table 7.20), generalized estimating equations (GEE) were used instead of MANOVA, as the latter requires at least $n > 30$ per group to have adequate power. In contrast, GEE provides consistent results even with small sample sizes. Fang (2019) proposed that for normally distributed response variables, GEE requires a minimum of 10 samples per group, while for binomial distributed variables, 20 samples per group are needed. Since the current data were predominantly normal, the present sample sizes were deemed adequate for GEE analysis. The GEE approach allowed for testing group differences despite having limited sample sizes that would be underpowered for methods like MANOVA. With mostly normal data and meeting the sample size criteria outlined by Fang (2019), GEE provided a consistent analytic approach suitable for the current data.

Table 7. 20 Test of Normality on Experiential Variables for T1, T2, and T3 Using Shapiro-Wilk Tests (n = 40) in Study 3

	Medium	***T1***			***T2***			***T3***		
		Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.
M_AffEase	Audio	.951	21	.349	/	/	/	/	/	/
	Avatar	.957	19	.519	/	/	/	/	/	/

M_RelEase	Audio	.929	21	.132	.924	21	.106	.939	21	.207
	Avatar	.903	19	.055	.901	19	.051	.943	19	.303
M_ValEase	Audio	.938	21	.203	.932	21	.150	.958	21	.475
	Avatar	.880	19	.021*	.958	21	.475	.928	19	.157
M_UseIntR	Audio	.807	21	<.001*	.898	21	.032	.985	21	.976
	Avatar	.918	19	.104	.896	19	.041*	.893	19	.036*
M_TrustP	Audio	.969	21	.709	.963	21	.585	.926	21	.116
	Avatar	.975	19	.876	.936	19	.219	.888	19	.030*
M_UseP	Audio	.846	21	.004*	.907	21	.047*	.901	21	.037*
	Avatar	.932	19	.186	.939	19	.258	930	19	.173

* This is a lower bound of true significance.

The analytic approach began with the application of GEEs to the complete dataset (*n* = 40) with no missing data. Two models were devised: **Model 1** analyzed effects of **Time**, **Medium**, and their interaction (**Time** × **Medium**) on the dependent variables; if non-significant, **Model 2** reassessed just **Time** and **Medium** main effects. **Depression** was added as covariates.

SPSS 28 GEE models (**Model 1** and **Model 2**) for each dependent variable included **Medium** as a fixed factor and **Time** as a within-factor. A gamma distribution with an identity link function and an unstructured working correlation matrix was used to account for skewness and different rates of change between time points, respectively.

Model 1 (with interactions) results are on the upper left side of Table 7.21. No significant interaction occurred. Thus, **Model 2** was run for variables. **Model 2** (without interactions) results are on the upper right side of Table 7.21. The main **Time** effect was significant on *M_UseIntR* ($\chi 2$ = 7.245, *df* = 2, *p* = .027)*, M_TrustP* ($\chi 2$ = 11.709, *df* = 2, *p* = .003) and *M_UseP* ($\chi 2$ = 8.873, *df* = 2, *p* = .012) in **Model 2**.

Within **Model 2**, higher depression scores significantly predicted lower relevance (*M_RelEase: B* = -0.005, *p* = .045) and valence (*M_ValEase: B* = -0.007, *p* < .001) of the robot for ease-of-use interaction, use intention of the robot (*M_UseIntR: B* = -0.007, *p* < .001), trust (*M_TrustP: B* = -0.006, *p* < .001), and usefulness evaluation (*M_UseP: B* = -0.006, *p* = .021) of the program, controlling for condition and time. The Wald 95% CI for the

depression coefficient was entirely negative, indicating robust negative relationships. These findings suggest that increased depressive symptoms are associated with the overall experience and perception of the social robot, even when accounting for experimental condition and changes over time.

GEE analyses were replicated without the participants reporting decreasing or low satisfaction trends. The significance remained the same except for the extract number (See bottom part of Table 7.21), indicating that the participants reporting decreasing or low satisfaction trend did not change the experience pattern among the participants. Therefore, pairwise comparison between **Time** among *M_UseIntR*, *M_TrustP* and *M_UseP* were only checked with the complete dataset.

Table 7. 21 Tests of Generalized Estimating Equation Model Effects on Experiential Variables (Model 1: With Interaction Effect and Model 2: With Interaction Effect) with $n = 40$ and $n = 33$ in Study 3

Experiential variables	**Model 1: Type III**			**Model 2: Type III**		
	Wald Chi-Square	*df*	Sig.	Wald Chi-Square	*df*	Sig.
		$N = 40$				
M_RelEase						
Medium	.242	1	.623	.027	1	.871
Time	.897	2	.638	.809	2	.667
Depression	3.737	1	.053	3.974	1	.046
Medium × Time	1.910	2	.385	/	/	/
M_ValEase						
Medium	.196	1	.658	.160	1	.690
Time	2.129	2	.345	2.036	2	.361
Depression	16.428	1	<.001	15.427	1	<.001
Medium × Time	.347	2	.841	/	/	/
M_UseIntR						
Medium	.043	1	.836	.042	1	.838
Time	7.584	2	.023	7.245	2	.027
Depression	15.222	1	<.001	14.856	1	<.001
Medium × Time	2.432	2	.296	/	/	/
M_TrustP						
Medium	.354	1	.552	.344	1	.557
Time	12.071	2	.002	11.709	2	.003
Depression	15.945	1	<.001	15.976	1	<.001
Medium × Time	.102	2	.950	/	/	/
M_UseP						
Medium	.963	1	.326	.937	1	.333

Time	8.791	2	.012	8.873	2	.012
Depression	5.408	1	.020	5.331	1	.021
Medium × Time	.083	2	.959	/	/	/
n (without satisfaction outliers) = 33						
M_RelEase						
Medium	.031	1	.861	.002	1	.961
Time	4.602	2	.100	3.519	2	.172
Depression	3.881	1	.049	4.035	1	.045
Medium × Time	2.508	2	.285	/	/	/
M_ValEase						
Medium	.085	1	770	.094	1	.760
Time	1.262	2	.532	1.149	2	.562
Depression	18.871	1	<.001	18.192	1	<.001
Medium × Time	1.274	2	.529.	/	/	/
M_UseIntR						
Medium	.804	1	.370	.551	1	.485
Time	13.134	2	.001	12.580	2	.002
Depression	11.221	1	<.001	15.861	1	<.001
Medium × Time	1.223	2	.543	/	/	/
M_TrustP						
Medium	.502	1	.479	.500	1	479
Time	11.876	2	.003	10.819	2	.004
Depression	19.838	1	<.001	19.836	1	<.001
Medium × Time	.047	2	.977	/	/	/
M_UseP						
Medium	.015	1	.902	.019	1	.889
Time	10.438	2	.005	9.858	2	.007
Depression	8.650	1	.003	8.439	1	.004
Medium × Time	.060	2	.971	/	/	/

Pairwise comparisons in Table 7.22 showed higher *M_UseIntR* ($MD_{(T3-T1)} = .204$, $p = .022$), *M_TrustP* ($MD_{(T3-T1)} = .330$, $p = .002$), and *M_UseP* ($MD_{(T3-T1)} = .279$, $p = .010$) scores in *T3* versus *T1* within **Model 1**. The *M_RelEase* and *M_ValEase* remains the same.

Table 7. 22 Pairwise Comparisons on M_UseIntP, M_TrustP, and M_UseP with Model 1 and Model 2 in Study 3

Condition	Means (SE)*	Mean Difference (*T3- T1*)	Std. Error	95% Wald Confidence Interval [Lower, Upper]	Sig
		M_UseIntP (within Model 1)			
T1	**4.369 (.122)**	.204	.076	-.386. -.021	.022
T3	**4.573 (.125)**				
		M_TrustP (within Model 1)			
T1	**4.259 (.092)**	.330	.098	-.566, -.094	.002
T3	**4.589 (.102)**				
		M_UseP (within Model 1)			
T1	**4.249 (.106)**	.279	.095	-.506, -.052	.010
T3	**4.527 (.126)**				

Note. The means reported for the GEE models represent model-estimated marginal means that account for correlations and other model parameters like outlier removal. In contrast, the ANOVA provides unweighted raw means. The GEE estimated means can more accurately reflect the true population average response.

7.3.6 Bayesian Analyses on Experiential Variables

Bayesian ANOVAs in JASP (van den Bergh et al., 2022) were conducted to examine the experience similarity across **Time** among participants guided by different robots. They also served as data analysis triangulation for the GEE results.

7.3.6.1 Bayesian Repeated Measures ANOVAs

A series of Bayesian repeated measures ANOVAs were conducted for each dependent variable with the between-subjects factor of **Medium** (*Audio* versus *Avatar*) and the within-subjects factor of **Time** across three timepoints (*T1*, *T2*, *T3*). **Depression** was not included as a covariate to avoid increasing the complexity of models given the limited sample size (n = 40), which could reduce the detectability of the main effects of interest.

The results in Table 7.23 showed that for robot relevance (*M_RelEase*) and valence (*M_ValEase*) for ease-of-use, the models including main effects of **Time** and **Medium** as well as their interaction provided the slight to strong evidence supporting the null hypotheses, suggesting similar between **Medium** over **Time** on these variables.

However, for robot use intention (*M_UseIntR*), program trust (*M_TrustP*), and program usefulness (*M_UseP*), models with only main effects of time or the combined model with main effects of time and condition showed slight to moderate evidence against the null. The data thus provide evidence for differences in these measures over time, regardless of **Medium**.

Table 7. 23 Bayesian Repeated Measures ANOVA Results for Experiential Variables with n = 40 in Study 3

Measure		Models				
		Null	**Medium**	**Time**	**Time + Medium**	**Time + Medium + Time × Medium**
M_RelEase	BF_{10}	1.000	0.359	0.128	0.047	0.014
	Evidence strength to H0	/	Slight	Moderate	Moderate	Strong
M_ValEase	BF_{10}	1.000	0.431	0.153	0.068	0.011
	Evidence strength to H0	/	Slight	Moderate	Moderate	Strong
M_UseIntR	BF_{10}	1.000	0.542	1.359	0.746	0.207
	Evidence strength to H0	/	Slight	Slight rejected	Slight	Moderate
M_TrustP	BF_{10}	1.000	0.438	5.778	2.712	0.356
	Evidence strength to H0	/	Slight	Moderate rejected	Slight rejected	Slight
M_UseP	BF_{10}	1.000	0.494	1.937	0.984	0.146
	Evidence strength to H0	/	Slight	Slight rejected	Slight	Moderate

Note. "Slight" for BF_{10} values that do not provide moderate evidence for H0 (greater than 0.334 but less than 1); "Moderate" for BF10 values between 0.334 and 0.033; "Strong" for BF10 values between 0.033 and 0.010; "Very Strong" for BF10 values less than 0.010.

7.3.6.2 Independent and Paired Comparison

Follow-up Bayesian paired comparisons (Table 7.24) were conducted on robot use intention (*M_UseIntR*), program trust (*M_TrustP*), and program usefulness (*M_UseP*) between *T1* and *T3*, given the significant time effects identified in the earlier GEE models. Bayesian Wilcoxon signed-rank tests provided Bayes factors (BF_{10}) quantifying evidence for differences between each timepoint comparison. There was very strong evidence for increases from *T1* to *T3* in *M_UseIntR, M_TrustP* and *M_UseP*, which corroborated the earlier GEE findings.

Table 7. 24 Bayesian Wilcoxon Signed-Rank Test (BF_{10} Value in Each Cell) with n = 40 in Study 3

Measure 1		Measure 2	BF_{10}	W	Rhat
M_UseIntR1	-	*M_UseIntR2*	0.782	202.500	1.003
M_UseIntR1	-	*M_UseIntR3*	16.532	99.000	1.003
M_UseIntR2	-	*M_UseIntR3*	0.550	177.000	1.001
M_TrustP1	-	*M_TrustP2*	0.317	198.500	1.000
M_TrustP1	-	*M_TrustP3*	45.099	112.000	1.023
M_TrustP2	-	*M_TrustP3*	0.898	132.000	1.005
M_UseP1	-	*M_UseP2*	0.600	156.500	1.001
M_UseP1	-	*M_UseP3*	28.956	115.500	1.006
M_UseP2	-	*M_UseP3*	0.626	76.000	1.001

7.3.7 Path Analyses on Experiential Variables

Path analyses were conducted at each of the three timepoints (*T1*, *T2*, *T3*) using SmartPLS for PLS-SEM to assess the impacts of experience and evaluation over time. Bootstrap methods estimated standardized path coefficients (β) and *p*-values (see Table 7.25). The *Pre_Exp* represented the evaluations of the last experience, including the robot use intention (*M_UseIntR*), on program usefulness (*M_UseP*) and trust evaluations (*M_TrustP*).

At *T1*, the path from *M_AffEas1e* to *M_RelEase1* was significant ($\beta = 0.385$, $p = .001$), but the path to *M_ValEase1* was not ($\beta = 0.206$, $p = .150$). *M_ValEase1* significantly predicted *M_UseIntR1* ($\beta = 0.428$, $p = .002$), but *M_RelEase1* did not ($\beta = 0.076$, $p = .631$). *M_UseIntR1* strongly predicted both *M_UseP1* ($\beta = 0.574$, $p < .001$) and *M_TrustP1* ($\beta = 0.615$, $p < .001$).

At *T2*, prior experience (*Pre_Exp2*) significantly predicted both *M_RelEase2* ($\beta = 0.378$, $p = .004$) and *M_ValEase2* ($\beta = 0.709$, $p < .001$). *M_ValEase2* again strongly predicted *M_UseIntR2* ($\beta = 0.757$, $p < .001$), while *M_RelEase2* did not ($\beta = 0.063$, $p = .670$). *M_UseIntR2* remained a significant predictor of *M_UseP2* ($\beta = 0.789$, $p < .001$) and *M_TrustP2* ($\beta = 0.564$, $p < .001$).

The *T3* results followed a similar pattern. *Pre_Exp3* predicted *M_RelEase3* ($\beta = 0.543$, $p < .001$) and *M_ValEase3* ($\beta = 0.704$, $p < .001$). *M_ValEase3* predicted *M_UseIntR3* ($\beta = 0.620$, $p < .001$), but not *M_RelEase3* ($\beta = 0.119$, $p = .501$). *M_UseIntR3* significantly predicted both *M_UseP3* ($\beta = 0.748$, $p < .001$) and *M_TrustP3* ($\beta = 0.679$, $p < .001$).

In summary, the path models demonstrate the perceived affordance predicted the valence of ease-of-use affordance but not the relevance of ease-of-use affordance at the first occurrence of the **eiIBM_RobotV2**. However, the previous evaluation of the robot and program did predict further valence and relevance of ease-of-use affordance. It remained

stable over time that valence of ease-of-use affordance predicted use intention of the robot and thus the usefulness and trust evaluation of the program.

Table 7. 25 Path Analyses Results on Experiential Variables for T1, T2 and T3 in Study 3

Path	*T1*			*T2*			*T3*		
	β	*t*	*p*	β	*t*	*p*	β	*t*	*p*
M_AffEase -> M_RelEase	0.385	3.361	.001	/	/	/	/	/	/
M_AffEase -> M_ValEase	0.206	1.438	.150	/	/	/	/	/	/
Pre_Exp -> M_RelEase	/	/	/	0.378	2.876	.004	0.543	4.728	.000
Pre_Exp -> M_ValEase	/	/	/	0.709	8.844	.000	0.704	9.163	.000
M_RelEase -> M_UseIntR	0.076	0.481	.631	0.063	0.427	.670	0.119	0.672	.501
M_ValEase -> M_UseIntR	0.428	3.141	.002	0.757	7.253	.000	0.620	3.756	.000
M_UseIntR -> M_UseP	0.574	4.014	.000	0.789	10.071	.000	0.748	9.512	.000
M_UseIntR -> M_TrustP	0.615	5.297	.000	0.564	3.664	.000	0.679	6.935	.000

Note. Pre_Exp is the construct consisting of the variables in respond phase, i.e., M_UseIntR, M_UseP, M_TrustP.

7.3.8 Effects of Medium on Intervention Outcome Over Time

The GEEs analyzed the effects of **TestTime**, **Medium** (*Audio*, *Avatar*), and their interaction on six outcome measures. I included experience outliers in the initial GEE analyses ($N = 44$) to examine whether they still benefited from the intervention, despite reporting not enjoying the experience or consistent experience. Results were shown in Table 7.26 and Table 7.27. Significant main effects of **TestTime** emerged across all measures (ps $< .001$), indicating symptom improvements over the course of the intervention regardless of **Medium**. **Medium** × **TestTime** interactions were also found on most outcome measures (*DS_MS*, *SRT_PT*, *SRT_NT*, *WSAP_NER*, and *WSAP_PMR*; ps ≤ 0.02), suggesting differing degrees of change depending on modality. However, the interaction was not significant for *SST_TNR* ($\chi^2 = 3.717$, $p = .156$).

For *DS_MS*, the *Audio* and *Avatar* groups showed pre-post decreases of 10.67 and 11.70 points versus stability in *Control* group (3.71 change), ps = .000. These corresponded to medium-large within-group effect sizes (Hedge's gs = 0.99 and 1.08). *Control* group did not

significantly change (p = .195). No between-group differences existed at *pretest* and *posttest* (ps > .205).

On *SST_TNR*, the *Avatar* group reduced scores .160 with medium effect sizes (Hedge's g = 0.27, p = .024) while the *Audio* (0.160 change, p = .290) and the *Control* groups were stable (0.063 change, p = .951). No *pretest* or *posttest* differences emerged (ps = 1.000).

For *SRT_PT, Audio* and *Avatar* groups showed pre-post declines of 5.52 and 7.19 points with medium-large effect sizes (Hedge's gs = 1.85 and 2.50), $p \leq 0.005$. The *Control* group was stable (0.61 change, p = 1.000). There were no pretest differences between groups (ps = 1.000). However, the *Audio* group differed from the *Control* group with a medium effect size (Hedge's g = 2.50, p = .021).

On *SRT_NT*, *Audio* and *Avatar* increased scores 6.61 and 5.19 with medium effect sizes (Hedge's gs = 0.67 and 0.63), ps < .009. The *Control* group changed little (1.11) but not significant (p = 1.000). No *pretest* or *posttest* differences among the three groups emerged (ps > 061).

For *WSAP_NER*, the *Audio* and *Avatar* groups showed pre-post declines with medium-large effect sizes (*Audio*: Hedge's g = 0.47; *Avatar*: Hedge's g = 0.40), while the *Control* group was stable (.027 change, p = 1.000). At *posttest*, the *Audio* and *Avatar* groups differed from the *Control* group with medium-large effect sizes (Hedge's gs = 0.59 and 0.44), ps < .034.

On *WSAP_PMR*, the *Audio* and *Avatar* groups reduced scores 0.215 and 0.189 with medium-large effect sizes (Hedge's gs = 0.59 and 0.53), ps = .000. The *Control* group worsened but not significantly (p = .119). At *posttest*, the *Control* group remained highest and different from the *Audio* and *Avatar* groups with a medium-large effect size (Hedge's gs = 0.57 and 0.46), ps < 0.007.

In summary, the intervention led to pre-post improvements with medium to large effect sizes versus stability in the *Control* group. Critically, no baseline differences existed at *pretest*. Some *posttest* differences emerged, but the overall patterns demonstrate the effectiveness of the intervention in reducing symptoms and biases regardless of *Audio* or *Avatar* medium.

Table 7. 26 Means, Standard Errors, Percentage Change and Effect Sizes on Intervention Outcome. N = 62 in Study 3

Measures	n	Mean (SE)		change [95 % CI]	Hedge's effect size [95 % CI]		
		pretest	posttest	pre-post	pre (versus Control)	pre-post (within)	post (versus Control)
DS_MS							
Audio		29.83 (1.818)	19.15 (2.454)	10.67 [3.53, 17.82]	0.18 [-0.28, 0.65]	0.99 [0.42, 1.55]	-0.46 [-1.01, 0.12]
Avatar		28.62 (1.816)	16.92 (2.399)	11.70 [6.82, 16.58]	0.07 [-0.39, 0.53]	1.08 [0.52, 1.64]	-0.68 [-1.24, -0.11]
Control		27.89 (1.571)	24.17 (1.698)	3.72 [-.68, 8.12]	…	…	…
SST_TNR							
Audio		.593 (.0530)	.416 (.0630)	.178 [-.045, .401]	-0.06 [-0.28, 0.15]	0.31 [0.10, 0.53]	-0.26 [-0.48, -0.04]
Avatar		.661 (.0529)	.501 (.0640)	.160 [.011, .308]	0.06 [-0.15, 0.28]	0.27 [0.06, 0.48]	-0.11 [-0.32, 0.11]
Control		.627 (.0520)	.565 (.0531)	.063 [-.036, .162]	…	…	…
SRT_PT							
Audio		24.83 (1.299)	30.35 (1.427)	-5.52 [-10.03, -1.02]	0.37 [-0.08, 0.82]	-1.85 [-2.29, -1.41]	2.50 [2.06, 2.94]
Avatar		22.48 (1.228)	19.67 (1.435)	-7.19 [-11.98, -2.40]	-0.56 [-1.01, -0.11]	-2.50 [-3.04, -1.96]	2.33 [1.88, 2.77]
Control		23.89 (1.407)	23.28 (1.688)	0.61 [-3.02, 4.24]	…	…	…
SRT_NT							
Audio		21.30 (1.299)	14.70 (1.565)	6.61 [2.75, 10.47]	0.04 [-0.39, 0.46]	0.67 [0.23, 1.10]	-0.73 [-1.16, -0.30]
Avatar		20.00 (1.127)	14.71 (1.139)	5.29 [2.85, 7.72]	-0.24 [-0.66, 0.19]	0.63 [0.19, 1.06]	-0.72 [-1.15, -0.29]
Control		21.11 (1.157)	20.00 (1.144)	1.11 [-2.48, 4.70]	…	…	…
WSAP_NER							
Audio		.588 (.047)	.374 (.043)	.213 [.089, .337]	-0.18 [-0.42, 0.07]	0.47 [0.24, 0.70]	-0.59 [-0.92, -0.26]
Avatar		.617 (.045)	.425 (.049)	.192 [.065, .319]	-0.11 [-0.34, 0.13]	0.40 [0.16, 0.63]	-0.44 [-0.76, -0.12]
Control		.661 (.034)	.634 (.047)	.027 [-.047, .101]	…	…	…
WSAP_PMR							
Audio		.447 (.039)	.233 (.041)	.215 [.111, .318]	-0.24 [-0.46, -0.02]	0.59 [0.36, 0.82]	-0.57 [-0.79, -0.34]
Avatar		.456 (.039)	.267 (.033)	.189 [.094, .283]	-0.21 [-0.43, 0.01]	0.53 [0.30, 0.75]	-0.46 [-0.68, -0.23]
Control		.525 (.024)	.440 (.037)	.085 [-.009, .179]	…	…	…

Table 7. 27 Statistical effects and comparisons between TestTime and Medium with N = 62 in Study 3

Indicators		Measures					
		DS_MS	SST_TNR	SRT_PT	SRT_NT	WSAP_NER	WSAP_PMR
GEE effects	TestTime	$\chi2 = 42.883$, $p < .001$	$\chi2 = 13.663$, $p < .001$	$\chi2 = 21.216$, $p < .001$	$\chi2 = 38.888$, $p < .001$	$\chi2 = 37.410$, $p < .001$	$\chi2 = 53.407$, $p < .001$
	Medium	$\chi2 = 2.377$, $p = .305$	$\chi2 = 2.264$, $p = .322$	$\chi2 = 5.136$, $p = .077$	$\chi2 = 5,220$, $p = .074$	$\chi2 = 10.917$, $p = .004$	$\chi2 = 14.187$, $p < .001$

		TestTime × Medium	$\chi2 = 12.577$, $p = .002$	$\chi2 = 3.717$, $p = .156$	$\chi2 = 15.880$, $p < .001$	$\chi2 = 13.328$, $p = .001$	$\chi2 = 22.168$, $p < .001$	$\chi2 = 13.893$, $p < .001$
p value from pairwise comparisons	pre (between)	Audio versus Avatar	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Avatar versus Control	1.000	1.000	1.000	1.000	1.000	1.000
	pre-post	Audio pre -> post	.000	.290	.005	.000	.000	.000
		Avatar pre-> post	.000	.024	.000	.000	.000	.000
		Control pre -> post	.195	.951	1.000	1.000	1.000	.119
	post (between)	Audio versus Avatar	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	1.000	1.000	.021	.191	.001	.003
		Avatar versus Control	.205	1.000	.059	.061	.034	.007

I then repeated the GEE analyses after removing outliers on each measure (*DS*, *SRT*, *WSAP*, *SRT*). The overall pattern of results (Table 7.28 and Table 7.29) remained similar to the analyses with outliers. The one difference was that the *Avatar* group no longer showed a significant pre-post change on *SST_TNR* ($p = .063$) after excluding outliers. The lack of change on *SST_TNR*, which comprises negative sentences, could be due to it also relating to attention bias, which the **eiIBM_RobotV2** did not address.

Table 7. 28 Means, standard errors, percentage change and effect sizes on Intervention Outcome with outliers excluded in Study 3

Measures	n	Mean (SE)		change [95 % CI]	Hedge's effect size [95 % CI]		
		pretest	posttest	pre-post	pre (versus Control)	pre-post (within)	post (versus Control)
DS_MS							
Audio		29.77 (1.900)	19.14 (2.464)	10.64 [3.43, 17.84]	0.25 [-0.22, 0.72]	0.99 [0.42, 1.56]	-0.49 [-1.06, 0.09]
Avatar		28.90 (1.885)	17.00 (2.417)	11.90 [7.05, 16.75]	0.17 [-0.29, 0.64]	1.10 [0.53, 1.67]	-0.69 [-1.26, -0.12]
Control		27.12 (1.462)	24.47 (1.771)	2.65 [-.69, 5.99]	…	…	…
SST_TNR							
Audio		.574 (.0519)	.397 (.0633)	.177 [-.058, .412]	-0.14 [-0.37, 0.10]	0.31 [0.09, 0.54]	-0.29 [-0.52, -0.06]
Avatar		.658 (.0588)	.499 (.0721)	.159 [-.004, .329]	0.01 [-0.22, 0.23]	0.24 [0.02, 0.46]	-0.10 [-0.32, 0.13]
Control		.652 (.0494)	.567 (.0562)	.085 [.004, .166]	…	…	…
SRT_PT							
Audio		24.89 (.874)	31.67 (1.676)	-6.78 [-11.46, -2.09]	0.36 [-0.24, 0.97]	-2.59 [-3.18, -2.00]	2.90 [2.31, 3.49]
Avatar		22.63 (1.593)	30.56 (1.591)	-7.94 [-13.62, -2.26]	-0.58 [-1.18, 0.03]	-2.90 [-3.49, -2.31]	2.51 [1.90, 3.11]
Control		24.00 (1.559)	23.75 (1.853)	.25 [-3.50, 4.00]	…	…	…
SRT_NT							
Audio		20.83 (1.348)	13.39 (1.761)	7.44 [3.11, 11.78]	0.08 [-0.41, 0.56]	0.94 [0.45, 1.43]	-1.03 [-1.52, -0.54]
Avatar		19.81 (1.282)	14.37 (1.293)	5.44 [2.34, 8.54]	-0.30 [-0.88, 0.28]	0.82 [0.33, 1.31]	-0.73 [-1.22, -0.23]
Control		20.62 (1.199)	19.56 (1.576)	1.06 [-2.60, 4.73]	…	…	…
WSAP_NER							

Audio	.552 (.048)	.332 (.040)	.220 [.080, .360]	-0.22 [-0.45, 0.02]	0.49 [0.25, 0.73]	-0.64 [-0.96, -0.32]
Avatar	.577 (.048)	.413 (.059)	.164 [.022, .306]	-0.18 [-0.42, 0.05]	0.31 [0.05, 0.57]	-0.37 [-0.69, -0.04]
Control	.659 (.035)	.613 (.048)	.046 [-.140, .262]	…	…	…
WSAP_PMR						
Audio	.431 (.041)	.219 (.044)	.212 [.094, .330]	-0.26 [-0.49, -0.02]	0.55 [0.31, 0.79]	-0.57 [-0.90, -0.25]
Avatar	.477 (.040)	.277 (.035)	.200 [.107, .293]	-0.12 [-0.35, 0.12]	0.55 [0.31, 0.79]	-0.39 [-0.72, -0.07]
Control	.518 (.026)	.433 (.041)	.085 [-.019, .188]	…	…	…

Table 7. 29 Statistical effects and comparisons between TestTime and Medium with outliers excluded in Study 3

Indicators			Measures					
			DS_MS	SST_TNR	SRT_PT	SRT_NT	WSAP_NER	WSAP_PMR
GEE effects		TestTime	$\chi2 = 42.360$, $p < .001$	$\chi2 = 13.199$, $p < .001$	$\chi2 = 24.821$, $p < .001$	$\chi2 = 32.885$, $p < .001$	$\chi2 = 30.091$, $p < .001$	$\chi2 = 45.032$, $p < .001$
		Medium	$\chi2 = 1.867$, $p = .393$	$\chi2 = 3.812$, $p = .149$	$\chi2 = 4.779$, $p = .092$	$\chi2 = 4,559$, $p = .102$	$\chi2 = 14.332$, $p < .001$	$\chi2 = 11.332$, $p = .003$
		TestTime × Medium	$\chi2 = 18.537$, $p < .001$	$\chi2 = 2.519$, $p = .284$	$\chi2 = 15.904$, $p < .001$	$\chi2 = 13.417$, $p = .001$	$\chi2 = 15.844$, $p < .001$	$\chi2 = 12.320$, $p = .002$
p value from pairwise comparisons	pre (between)	Audio versus Avatar	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	1.000	1.000	1.000	1.000	1.000	1.000
		Avatar versus Control	1.000	1.000	1.000	1.000	1.000	1.000
	pre-post	Audio pre -> post	.000	.407	.000	.000	.000	.000
		Avatar pre-> post	.025	.063	.001	.000	.011	.000
		Control pre -> post	.299	.030	1.000	1.000	.972	.240
	post (between)	Audio versus Avatar	1.000	1.000	1.000	1.000	1.000	1.000
		Audio versus Control	1.000	.683	0.23	.135	.000	.005
		Avatar versus Control	.190	1.000	0.79	.164	.133	.053

7.3.9 Effects of Medium on Intervention Outcome Differences Over Time

Residual change scores of the experiential variables were calculated to analyze indicator change differences between **Medium**. *RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, *RES_NT* and *RES_PT* were obtained as the residual change scores of *DS_MS*, *SST_TNR*, *WSAP_NER*, *WSAP_PMR*, *SRT_NT*, and *SRT_PT* between *pretest* and *posttest* respectively.

With all participants ($N = 62$), a test of normality on the residual change across Medium showed that the data were normally distributed, except for *RSE_PT* in the Control group. Therefore, one-way MANOVA analyses on the residual change scores with Medium

(*Audio* versus *Avatar* versus *Control*) as dependent variables were conducted to understand the intervention effect difference between groups. Depression was added as a covariate to control its potential distortion on the intervention effect, given its high correlation with experience and its potential influence on intervention outcome, as observed in Study 1. Results showed no difference in the composite dependent variable between **Medium**, F (6, 36) = 1.109, p = .376, Pillai's V = .156, partial $\eta 2$ = .156. Follow-up one-way ANOVAs indicated no significant differences between Medium on all measures (all ps > .05). Depression did not affect the results. These results suggest that participants benefited from the **eiIBM_RobotV2** in symptoms of depression and negative interpretation biases, regardless of the type of robot guiding the **eiIBM** exercise. The results remained the same when engagement outliers, imagery outliers, and both were removed.

7.3.10 Effect of Experience on Intervention Outcome Difference

Hierarchical cluster analysis (HCA) using Ward's method with squared Euclidean distance was conducted to categorize participants based on experience ratings across 16 measures over three times. Inspecting the dendrogram and agglomeration schedule indicated a 3-cluster solution was optimal. The *Control* group was excluded from the following analyses as they did not have experience data.

The 3 experience clusters (*ExpRank*) were labeled 'high experience' (*Exp_H*, n =15), 'medium experience' (*Exp_M*, n = 23), and 'low experience' (*Exp_L*, n =5). The three participants clustered into low experience were the experience outliers detected beforehand. Due to the limited *Exp_L* sample size, the experiential outliers and *Exp_L* groups, a total of 6 participants, were not included in intervention outcome difference analyses.

Almost all residual depressive measures change scores were normally distributed within their *ExpRank* (except *Exp_H*'s *RES_PMR*: p < .001). A one-way MANOVA assessed

ExpRank (*Medium* versus *High*) differences in intervention outcomes. The composite dependent variable was significantly affected by *ExpRank*, F (6, 30) = .831, p = .555, Pillai's V = .143, partial $\eta 2$ = .143. Follow-up one-way ANOVAs showed significant differences between *ExpRank* on all experience measures (all *ps* <.001), except for *M_AffEase1*: F (1, 35) = .023, p = .881, partial $\eta 2$ = .001 and *M_RelEase1*: F (1, 35) = 2.45, p = .126, partial $\eta 2$ = .066. This indicated that all the participants with high or medium experience perceived the ease-of-use affordance and its relevance equally. Table 7.30 indicated that the *Exp_H* cluster had higher average ratings than that of the *Exp_M*, all *ps* <.05.

Table 7. 30 Means on Experience Measure across ExpRank in Study 3

	T1			***T2***			***T3***		
Measures	Exp_L	Exp_M	Exp_H	Exp_L	Exp_M	Exp_H	Exp_L	Exp_M	Exp_H
M_AffEase	4.00 (1.09)	4.59 (.72)	4.68 (.31)	/	/	/	/	/	/
M_RelEase	3.4 (.89)	4.20 (.94)	4.60 (.71)	3.80 (1.30)	4.09 (.90)	5.07 (.80)	3.70 (1.10)	4.17 (.76)	5.13 (.81)
M_ValEase	2.45 (.99)	4.22 (.62)	4.77 (.45)	3.30 (1.08)	3.96 (.54)	5.02 (.70)	2.35 (1.10)	4.20 (.51)	5.10 (.72)
M_UseIntR	2.56 (.91)	3.96 (.40)	5.13 (.72)	2.48 (1.06)	4.03 (.40)	5.33 (.61)	2.92 (.72)	4.25 (.37)	5.33 (.64)
M_TrustP	3.00 (.98)	3.87 (.56)	4.87 (.38)	3.44 (1.65)	3.94 (.56)	4.87 (.75)	3.36 (1.12)	4.24 (.47)	5.16 (.60)
M_UseP	1.87 (.77)	4.06 (.60)	4.69 (.51)	2.60 (.83)	4.00 (.46)	5.18 (.62)	2.33 (.47)	4.18 (.52)	5.20 (.65)
Num	5	23	15	5	23	15	5 23	15	5

A one-way MANOVA with *ExpRank* as independent variable was conducted to examine the intervention effect difference. No significant differences were found on residual change scores of evaluations (Check Table 7.31).

Table 7. 31 Univariate Test of ExpRank on Experiential residual change with N = 44 in Study 3

Effect	*F*	Hypothesis df	Error df	*MD(High-Medium)*	Sig.	Partial Eta Squared
RES_TNR	.084	1	35	-.026	.773	.002
RES_NER	.001	1	35	-.002	.977	.000
RES_PMR	1.697	1	35	.068	.201	.046
RES_NT	.569	1	35	1.715	.456	.016
RES_PT	.044	1	35	.490	.834	.001
RES_DS	.090	1	35	-1.20	.767	.003

7.3.11 Exploration Analyses

In Study 1, I did not find an association between depression and experience variables/overall experience rank. The non-correlation remained the same with *Audio* and *Video* participants only, r = -.088, p = .627, n = 33 (overall experience rank). However, in Study 3, depression showed a medium relation with overall experience rank, especially a strong correlation with experience variables at *T2* and *T3*. Specifically, more severe depression was associated with poorer participant experience. Despite this, overall experience rank did not influence therapy outcomes in present study, indicating that depression did not impact therapy outcomes despite affecting experience. This contrasts with Study 1 results where overall experience ranks affected therapy change.

To further examine the comparability and differences between **eiIBM_RobotV1** and **eiIBM_RobotV2** on experience and the cognitive bias reduction, I compared four groups: *Audio* and *Video* from **eiIBM_RobotV1**, and *Audio* and *Avatar* from **eiIBM_RobotV2**. The *Text* group from **eiIBM_RobotV1** had nonsignificant but descriptively lower experience than the *Audio* and *Video*, and chatbots were not the target medium in this study; thus, it was excluded from the analyses. Comparing these four groups allowed me to elucidate the impacts of exercise format and medium on user experiences and outcomes.

7.3.11.1 Depression Distribution between Study 1 and Study 3

An independent sample *t*-test was conducted to compare depression severity between Study 1 (N = 33, only audio bot and telepresence robot groups) and Study 3 (N = 44). Study 3 had a slightly higher percentage of participants in the mild depression category (6.8% versus 0%), and a lower percentage in the severe depression category (48% versus 52%) compared to Study 1. However, the average depression score was similar between Study 1 (M=30.27,

SD=9.56) and Study 3 (*M*=29.25, *SD*=8.65), *t* (75) =.43, *p* = .662. No significant differences emerged for mild, moderate, or severe depression categories.

The results (Table 7.32) indicate participants' depression levels were more sensitive to the exercise format in Study 3. One possible explanation might be the half-hour speaking and generation was tiring for the severely depressed participants.

Table 7. 32 Depression distribution on DepressLev in Study 1 and Study 3

Depression Level (*DepressLev*)	**Depression Score**				**Independent Samples Test**		
	Study 1		Study 3		t	df	Two-Side p
	Mean (SD.)	N (%)	Mean (SD.)	N (%)			
1	0	0 (/)	12.67 (1.155)	3 (6.8%)	/	/	/
2	16.20 (1.789)	5 (15%)	20.00 (/)	1 (2.2%)	/	/	/
3	26.64 (4.478)	11 (33%)	25.05 (2.272)	19 (43%)	1.291	28	.207
4	36.76 (7.336)	17 (52%)	36.10 (6.465)	21 (48%)	.299	36	.767
Total	30.27 (9.563)	33	29.36 (8.573)	44	.438	75	.662

7.3.11.2 Different Means of Intervention Outcomes Over Study 1 and Study 2

To exclude the possibility that differences in initial depression levels or cognitive biases explained effects on therapy outcomes, two-way MANOVAs were conducted with **Medium** (*Audio* versus *Video*) and **Exercise** (**eiIBM_RobotV1** versus **eiIBM_RobotV2**) as factors. The **eiIBM_RobotV1** utilized a telepresence robot with a video screen, while the **eiIBM_RobotV2** used an avatar with a virtual human face. Though not a live video feed, the avatar was an enhanced, interactive form of video communication. Thus, both the telepresence robot and virtual avatar were conceptualized as types of *Video* medium for comparison with *Audio*.

Dependent variables were initial depression (*DS_MC_1*) and negative interpretation bias (*SST_TNR_1*, *WSAP_NER_1*, *WSAP_PMR_1*, *SRT_PT_1*, and *SRT_NT_1*) indicators. Results showed no significant multivariate main or interaction effects, indicating no differences between conditions on initial depressive symptoms or cognitive biases.

A second two-way MANOVA examined the reduction in depressive symptoms or cognitive biases after intervention using residual change scores (*RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, *RES_NT* and *RES_PT*). Results showed no significant multivariate main or interaction effects. The only significance was found on the Univariate Test for *RES_PT* between *Video* groups, F (1, 73) = 4.375, p = .040, partial $\eta 2$ = .057. Participants reported higher *RES_PT* after the Avatar-guided **eiIBM_RobotV2** (M = 2.15, SE = 1.38) than the telepresence robot-guided **eiIBM_RobotV1** (M = -2.23, SE = 1.57). This indicates that after avatar interaction in **eiIBM_RobotV1**, participants were more likely to agree with positive interpretations of ambiguous scenarios.

When rerunning a series of two-way ANOVAs on residual change scores without outliers respectively, the significant difference in *RES_PT* between *Video* groups remained, F (1, 60) = 5.408, p = .023, partial $\eta 2$ = .083, while no significant differences emerged for other variables. *Video* group participants reported 5.133 higher on *RES_PT* after **eiIBM_RobotV2** exercise (M = 3.36, SE = 1.51) than that in **eiIBM_RobotV1** (M = -1.77, SE = 1.61) exercise**.**

7.3.11.3 Effect of Depression Rank on Experience Rank

A Chi-square test of independence was conducted to examine the relation between depression level (*DepressLev*) and experience rank (*ExpRank*) with Study 3 (Table 7.33). *Depression level* included four categories: none, mild, moderate, and severe. Experience rank included three categories: high (1), medium (2), and low (3).

The results showed a statistically significant association between depression level and experience rank, $\chi^2(6)$ = 15.33, p = .014. Participants with moderate to severe depression were more likely to be in the low experience rank group compared to those with no or mild depression. Of the 15 participants in the low experience rank group, 10 had moderate to

severe depression, while only 5 out of 29 participants in the medium to high experience rank groups had moderate to severe depression.

The significant linear-by-linear association, $\chi^2(1) = 8.51$, $p = .003$, further suggested that as depression level increased, experience rank tended to decrease. These findings indicate an association between higher depression and poorer user experience - participants with more severe depression symptoms were more likely to have lower quality experiences.

Table 7. 33 ExpRank*DepressLev Crosstabulation with N = 44 in Study 3

Expression Rank (*ExpRank*)	Depression Level (*DepressLev*)				
	1	2	3	4	Total
1	3	0	10	2	15
2	0	1	7	16	24
3	0	0	2	3	5
Total	3	1	19	21	44

A replication analysis using the Study 1 dataset (See Table 7.34) did not indicate a statistically significant association between depression level and experience rank $\chi^2(4) = 4.416$, $p = .397$. The linear-by-linear association was also not significant, $\chi^2(1) = 0.123$, $p = .782$, confirming the lack of an association in Study 1.

Overall, while Study 3 demonstrated a relation between higher depression and lower experience rank, this association was not replicated in Study 1.

Table 7. 34 ExpRank*DepressLev Crosstabulation with n = 40 in Study 3

Expression Rank (*ExpRank*)	Depression Level (*DepressLev*)				
	1	2	3	4	Total
1	0	1	6	4	11
2	0	1	3	5	9
3	0	3	2	8	13
Total	0	5	11	17	33

7.3.11.4 Different Means of Experience Over Study 1 and Study 3

A third two-way MANOVA examined experience after the first interaction, with **Medium** (*Audio* versus *Video*) and **Exercise** (**eiIBM_RobotV1** vs **eiIBM_RobotV2**) as

independent variables. The dependent variables were ease-of-use affordance (*M_AffEase1*), relevance (*M_RelEase1*), valence (*M_ValEase1*), and use intention (*M_UseInt1*). Noted that that I measured program use intention in Study 1, but robot use intention in Study 3 due to program use intention most contributed by emotional distress relief. Careful interpretation would be made of this comparison. A significant main effect of **Exercise** was found on *M_AffEase1*, F (1, 73) = 11.90, $p < .001$, partial $\eta 2 = .14$ and *M_ValEase1*, F (1, 73) = 4.64, $p = .035$, partial $\eta 2 = .06$. Participants reported higher ease-of-use affordance and valence in **eiIBM_RobotV1** (*M_AffEase1*:*M* = 5.12, *SE* = 0.12; *M_ValEase1*: *M* = 4.60, *SE* = 0.15) compared to **eiIBM_RobotV2** (*M_AffEase1*: *M* = 4.56, *SE* = 0.11; *M_ValEase1*: *M* = 4.18, *SE* = 0.13).

After removing experiential outliers, only the main effect of **Exercise** on *M_AffEase1* remained significant, F (1, 65) = 6.88, $p = .011$, partial $\eta 2 = .096$. Participants reported higher ease-of-use affordance in the **eiIBM_RobotV1** (*M* = 5.04, *SE* = 0.12) compared to the **eiIBM_RobotV2** (*M* = 4.64, *SE* = 0.10). Without outliers, a significant effect emerged for usefulness of the program (*M_UseP1*) between **eiIBM_RobotV1** (*M* = 4.71, *SE* = 0.197) and **eiIBM_RobotV2** (*M* = 4.17, *SE* = 0.169) when guided by *Video*, F (1, 65) = 4.33, $p = .041$, partial $\eta 2 = .062$.

A fourth two-way MANOVA on experience variables after the third interaction (*T2*) revealed significant main effects of **Exercise** on *M_ValEase2* F (1, 73) = 5.57, $p = .021$, partial $\eta 2 = .07$, and *M_UseP2*, F (1, 73) = 5.77, $p = .019$, partial $\eta 2 = .07$. Ease-of-use valence and perceived usefulness were rated higher in the **eiIBM_RobotV1** (*M*s = 4.69 and 4.74, *SE*s = 0.15 and 0.16) compared to the **eiIBM_RobotV2** (*M*s = 4.24 and 4.24, *SE*s = 0.13 and 0.14). However, these effects disappeared when outliers were removed.

A fifth two-way MANOVA on experience after the fifth interaction were examined using the same variables in *T3*. Results indicated a multivariate main effect of **Exercise**, F (4,

70) = 2.54, p = .047, partial $\eta 2$ = .13, which was reduced to a trend after excluding outliers, F (4, 62) = 2.46, p = .055, partial $\eta 2$ = .14.

Univariate tests showed a significant effect of **Exercise** on *M_UseIntR3*, F (1, 65) = 4.00, p = .050, partial $\eta 2$ = .06, with lower use intention in **eiIBM_RobotV1** (M = 4.05, SE = 0.21) compared to **eiIBM_RobotV2** (M = 4.47, SE = 0.21). A significant interaction between **Medium** and **Exercise** was also found for *M_UseIntR3*, F (1, 65) = 4.21, p = .044, partial $\eta 2$ = .06, with lower use intention in **eiIBM_RobotV1** versus **eiIBM_RobotV2** for the *Audio* medium.

7.3.11.5 Summary of Different Means of Experience and Cognitive Outcome Over Study1 and Study 3

In summary, MANOVAs showed no baseline differences between conditions on depression or cognitive biases, indicating comparable starting points. Follow-up analyses revealed no differences in residual symptom change scores, except on positive elaboration interpretation endorsement (*RES_PT*). Video-guided participants reported greater increases in *RES_PT* after **eiIBM_RobotV2** versus **eiIBM_RobotV1**, suggesting the avatar's improved naturalness and synchronicity solicited more willingness to consider and endorse positive interpretations.

Effects emerged on user experience variables. After the first interaction, participants reported higher ease-of-use affordance and valence in **eiIBM_RobotV1**, likely due to **eiIBM_RobotV2**'s initial difficulty for some participants. By the third interaction, ease-of-use valence and perceived usefulness were rated higher in **eiIBM_RobotV1**, but these advantages disappeared after removing outliers. By the fifth interaction, use intention was lower for **eiIBM_RobotV1** versus **eiIBM_RobotV2** in *Audio* conditions only.

7.3.11.6 Trust Base

Three multiple response sets assessed participants' trust bases at three timepoints during the human-robot interactions (*T1*, *T2*, *T3*). Each set included five dichotomous variables: robot's capabilities, emotional experience with robot, past robot experiences, understanding of robot capabilities, and overall view of robots.

Frequencies (Table 7.35) showed robot's capabilities as the most frequently selected trust base across timepoints, endorsed by 72.1% to 75.0% of participants. Overall view of robots was also commonly selected (45.5% to 61.4% of cases). Emotional experience, past interactions, and understanding were endorsed by 30.2% to 59.1% of participants, with experience (both previous and current) being increasingly endorsed over time.

Table 7. 35 TrustBase criteria Frequencies in Study 3

	T1			***T2***			***T3***		
TrustBase	N	Percent	Percent of Cases	N	Percent	Percent of Cases	N	Percent	Percent of Cases
Zora's capability	31	31.6%	72.1%	32	30.8%	72.7%	33	27.0%	75.0%
Emotional experience with Zora	16	16. 3%	37.2%	18	17.3%	40.9%	26	21.3%	59.1%
Past robot experiences robot	13	13.3%	30.2%	15	14.4%	34.1%	21	17.2%	47.7%
Past robot understanding	13	13.3%	30.2%	19	18.3%	43.2%	15	12.3%	34.1%
Overall view of robots	25	25.5%	58.1%	20	19.2%	45.5%	27	22.1%	61.4%
Num	98	100%	227.9%	104	100%	236.4%	122	100%	277.3%

Crosstabulations (Table 7.36) examined trust bases by depression level (medium versus severe). Those with severe depression more frequently selected emotional experience across timepoints (21.4% to 22.6% vs 10.9% to 20.3%), while the medium depression group showed slightly higher frequencies of selecting past experiences (9.5% to 17.0% vs 15.2% to 17.8%) and understanding (9.5% to 11.3% vs 13.6% to 20.0%). The trust of severely depressed participants was more affected by their current experience compared to those with medium depression.

Table 7. 36 TrustBase*DepressLev crosstabulation in Study 3

Measures	*DepressLev*	*T1*		*T2*		*T3*	
		N	% within *DepressLev*	N	% within *DepressLev*	N	% within *DepressLev*
Zora's capability	3	13	28.3%	12	26.7%	13	22.0%
	4	15	35.7%	16	32.7%	18	34.0%
Emotional experience with Zora	3	5	10.9%	7	15.6%	12	20.3%
	4	9	21.4%	9	18.4%	12	22.6%
Past experiences of robot	3	8	17.4%	8	17.8%	10	16.9%
	4	4	9.5%	7	14.3%	9	17.0%
Past understanding of robot	3	7	15.2%	9	29.0%	8	13.6%
	4	4	9.5%	9	18.4%	6	11.3%
Overall view of robots	3	13	28.3%	9	20.0%	16	27.1%
	4	10	23.8%	9	16.3%	8	15.1%

Note. number of participants with 3rd (Medium) depression and with 4th (Severe) depression was 19 and 21.

7.5 Summary and Discussion

This study replicated the experiment in Study 1 with the variant of **eiIBM_RobotV1-eiIBM_RobotV2**. The **eiIBM_RobotV2** program was designed to be delivered by the virtual avatar and the audio bot. It explored user's experience with **eiIBM_RobotV2** and examined the experience impact on intervention outcomes. The study addressed three questions:

RQ3.1 Are user perceptions largely similar between virtual avatars and audio bots delivering the **eiIBM_RobotV2** exercise?

RQ3.2 Given the improved **eiIBM_RobotV2** exercise, do the intervention outcomes differ depending on whether it is delivered by virtual avatars or audio bots?

RQ3.3 Dose **eiIBM_RobotV2** produce higher or at least comparable intervention effects on negative interpretation biases and depressive symptoms compared to **eiIBM_RobotV1**?

7.5.1 Discussion on Research Questions

To address the **RQ3.1**, GEE models with **Medium** (*Avatar* versus *Audio*) and **Time** (*T1*, *T2*, *T3*) as factors showed no significant effects of **Medium** or the interaction between **Medium** and **Time**, except for **Time** effect on the improvement of use intention of robot, program usefulness and program trust in *T3* compared *T1*. Bayesian ANOVAs further corroborated the lack of differences between conditions over time. The path analysis with *T1*

data revealed that the initial ease-of-use affordance positively predicted the valence but not the relevance in the first session. These findings supported hypothesis **H6.1** that avatars and audio bots would elicit equal initial ease-of-use affordance between the two media, and partially supported **H6.2**, which predicted equal valence but not the effect on relevance from perceived affordances.

Regarding **H6.3** that avatars and audio bots would produce comparable robot use intentions, GEEs and Bayesian tests again showed no significant differences between conditions on this variable over time. Path analysis further revealed that it was valence rather than relevance of robot contributing to the equal use intention of the medium, partially supported the **H6.3**. For **H6.4** that avatars would increase program usefulness and trust perceptions, path analyses showed that use intention consistently predicted usefulness and trust regardless of medium. No direct comparative evidence emerged to support avatars uniquely improving evaluations. Regarding **H6.5**, sustained exposure would maintain experience similarity, GEEs indicated no significant **Medium** x **Time** interactions and path models demonstrated stable patterns, aligning with the hypothesis.

To evaluate **RQ3.2**, GEE models examined the effects of **TestTime** (*pretest*/*posttest*), **Medium** (*Avatar* versus *Audio*), and their interaction on cognitive outcome measures of depression, medium negative interpretation bias, elaborative interpretation bias, and automatic interpretation bias. Results demonstrated significant improvements from pre to post-test across all outcome measures, regardless of **Medium**. No baseline differences existed between conditions at pretest. While some **Medium** by **TestTime** interactions emerged, the overall patterns indicated comparable benefits in reducing symptoms and cognitive biases after completing the **eiIBM_RobotV2** exercise, irrespective of virtual avatar or audio bot guidance.

These findings provided support for hypothesis **H6.6**, which predicted equal intervention effectiveness between the two media given the same underlying exercise mechanism. When examining changes in residualized outcome scores using MANOVA, there were also no significant differences found between media on changes in depression, biases, or most cognitive measures. This further substantiated the effectiveness of **eiIBM_RobotV2** in improving outcomes for both virtual avatar and audio-bot conditions.

To address **RQ3.3**, two-way MANOVAs compared starting points, outcome changes, and experience variables between Study 1 (**eiIBM_RobotV1**) and Study 3 (**eiIBM_RobotV2**). Results indicated no differences between conditions at baseline on depression levels or cognitive biases, suggesting comparable starting points. Follow-up analyses also showed no differences in revisualized changes on most outcome measures after completing the intervention.

The one exception was that participants in the video-guided condition (using either a telepresence robot in Study 1 or virtual avatar in Study 3) reported greater pre-post increases in positive interpretation endorsement after completing **eiIBM_RobotV2** compared to **eiIBM_RobotV1**. This provides partial support for hypothesis **H6.6**, suggesting the improved avatar medium specifically helped increase positive interpretation tendencies. No other significant between-group differences emerged for either video or audio conditions.

Though higher depression was related to lower experience in the present study, the overall experience did not influence the results. Therefore, **H6.7** and **H6.8** were rejected.

7.5.2 Theoretical and Practical Implications

7.5.2.1 Robot Relevance Not Predicting Use Intention and Not Predicted by Affordance

In the I-PEFiC framework, relevance and valence represent two key dimensions for comparing perceived affordances to concerns and forming engagement intention (Hoorn.

2012). However, the framework does not necessitate that relevance and valence will equally contribute to intentions.

The current findings showed robot relevance did not significantly predict robot use intention or steam from perceived ease-of-use affordance, in contrast to Study 2 using **eiIBM_RobotV1** where both relevance and valence contributed to robot use intention in path analyses. This discrepancy may involve differences in the affordance content between programs. **eiIBM_RobotV1** provided exercise guidance, meeting users' values for concentration and emphasizing efficient interaction. Conversely, **eiIBM_RobotV2** also encouraged and motivated participants through natural conversation, potentially enhancing affordance content towards perceiving natural, easy interactions over efficiency. With this affordance focus, valence (e.g., "*The experience with [name of robot], I felt it is pleasant*" was more salient than relevance (e.g., "*In terms of providing help, [name of robot] make me complete tasks faster*") for use intentions.

7.5.2.2 Depression Severity Effects on Experience

In the present study using **eiIBM_RobotV2**, participants with more severe depression reported poorer quality experiences compared to those with mild or moderate levels of depression. This could be explained by emotional reactivity and reward processing biases associated with depression (Bishop & Gagne, 2018; Russo & Nestler, 2013; Kupferberg et al., 2016; Forbes & Dahl, 2011; Silk et al., 2021), where higher severity relates to greater reactivity to negative events and reduced reward responsiveness to positive effects (Seepersad, 2014). These affective processing biases could make interaction more effortful and less rewarding for those with severe depression, reducing user experience quality. Additionally, depressive rumination (Nolen-Hoeksema, 1991) could further diminish

experience quality if attention wanders away from the interaction and becomes inwardly focused on negative thoughts, competing with outward engagement.

Interestingly, this trend was absent in Study 2 using **eiIBM_RobotV1**, where some participants reported not feeling oppressed when the robot promoted positive resolutions they disagreed with, recognizing its non-human status. Conversely, **eiIBM_RobotV2**'s more humanlike natural speech, response understanding, and lack of visual robotic identity cues may have elicited a greater sense of affective interaction. As Zhang et al. (2021) found, social rewards during human-robot interactions are less impacted by depression than in human-human interactions. If robots lack sufficient cues denoting their non-human status, the limitations of human-human interaction may manifest.

Researchers and designers developing therapeutic robots should carefully examine the consequences of humanlike-ness. Excessively anthropomorphic traits could elicit problematic biases in depressed users, deteriorating their experience. Identifying an optimal balance of humanlike features while preserving robotic identity may circumvent the detrimental effects of severe depression.

7.5.2.3 Experience Not Affecting Intervention Outcomes

The finding that overall experience quality (high/medium/low ranks) did not impact cognitive outcome changes suggests that participants can benefit from the **eiIBM_RobotV2** intervention even with a suboptimal user experience. Several “low experience” outliers exhibited cognitive improvements comparable to or exceeding “high experience” participants, indicating that poorer experience does not preclude gaining cognitive restructuring skills.

One explanation involved the potency of the core **eiIBM** ingredients. As a structured CBT technique, repetitive practice generating positive resolutions may improve thinking

pattern irrespective of delivery quality. Meta-analyses indicate that the magnitude of early cognitive gains independently predicts later symptom changes, regardless of the therapeutic relationship (Forand & DeRubeis, 2013). For skills-based CBT, directly completing the key exercises could be sufficient to increase cognitive flexibility and clinical improvement (Karver et al., 2006).

In **eiIBM_RobotV2**, the central repetition of generating alternative interpretations benefited even participants who struggled with engagement or imagery. This active ingredient seemed to overcome experience deficiencies and improve cognitive outcomes. In contrast, **eiIBM_RobotV1**'s passive design made experience more crucial; poor experiences likely hindered engagement, limiting exercise completion and gains. **eiIBM_RobotV2** participants had to actively generate resolutions themselves, ensuring exercise completion and cognitive benefits despite dissatisfaction. Although **eiIBM_RobotV1** participants also had to provide responses, its passive nature allowed disengagement that may have reduced intervention effects for dissatisfied users. The active participation compulsory in **eiIBM_RobotV2** appeared to overcome experience deficiencies through forced engagement. This highlights how the degree of involvement required, whether passive or active, impacts the relevance of user experience on therapeutic outcomes. The **eiIBM_RobotV2** findings suggest that compelled active participation can produce cognitive gains irrespective of experience, while more passive interventions may rely on experience to motivate engagement and completion.

These findings suggest structured CBT interventions' core ingredients directly improve outcomes, with experience providing incremental benefits. However, researchers should remain vigilant about the mental health impacts of negative experiences on vulnerable users (O'Grady et al., 2010; Steger & Kashdan, 2009). Although **eiIBM_RobotV2** participants completed exercises despite dissatisfaction, unwilling participants may avoid exercises or drop out when self-administered outside experiments. While structured CBT

exercises may produce cognitive gains irrespective of experience, maximizing acceptability and engagement could amplify and extend benefits (Glenn et al., 2013).

7.5.2.4 Improving Experience (i.e., Use Intention, Usefulness and Trust) Over Time

With **eiIBM_RobotV2**, results showed user experience, in terms of Use Intention, Usefulness and Trust, improved across sessions. The robot frequently encouraged and affirmed participants as compensation for the demanding exercise format. For those who perceived the robot as different from humans, they could be pushed without eliciting resistance.

Prior research has found that trust in social robots increases with sustained interactions as users calibrate their expectations and build confidence (Yang et al., 2017). This concept can be applied to conversational agents as well. Gaver (1991) suggests that as users become more familiar with a technology, its affordances become more obvious. In the context of conversational agents, this implies that an agent's affordances may become more apparent to users over time as they gain familiarity with the system through repeated interactions. Furthermore, successfully generating one's own positive resolutions likely raises self-efficacy, enhancing motivation and engagement. Hoorn (2020b) proposes that friendship building occurs over time, indicating that the relationship between the user and the conversational agent might strengthen as the user engages in more interactions.

Allowing participants to self-generate resolutions addressed autonomy needs (Hoppitt et al., 2010; Schweizer et al., 2011), which improves motivation in self-determination theory (Deci & Ryan, 1985). The sense of authorship from creating personalized positive endings, versus the system imposing them, increased perceived relevance of the scenarios over sessions. Additionally, successfully applying new thinking patterns in an increasing number

of real-life situations helped demonstrate the exercise's usefulness. The findings highlighted the long-term experience advantage of **eiIBM_RobotV2.**

7.5.2.5 Initial Poorer Experience in eiIBM_RobotV2 than eiIBM_RobotV1

Initial comparisons of the interaction's perceived ease-of-use revealed lower scores after participants' first exposure to **eiIBM_RobotV2** versus **eiIBM_RobotV1**. Some participants expressed uncertainty initially regarding the new **eiIBM_RobotV2** format requiring self-generated responses. The complex structure and need for active participation imposed greater cognitive demands, which were more difficult, especially for those with depression. However, over repeated sessions, this experience discrepancy disappeared as ease-of-use ratings in **eiIBM_RobotV2** caught up to match or exceed **eiIBM_RobotV1** levels.

This pattern aligns with prior findings that initial difficulty from open-ended dialogue can encourage closer engagement to master the task, eventually enhancing learning and motivation (Rossi et al., 2022). Sustained exposure to conversational agents has been shown to increase perceived ease-of-use over time as familiarity develops, which boosts intentions (Fan et al., 2020). Additionally, accumulated experience of affordances can shape subsequent perceptions (Hoorn, 2020b).

Theoretically, these findings highlight the importance of examining temporal dynamics in user experience rather than just initial impressions. While prior knowledge and challenge may impede early interactions, sustained practice and mastery can transform initial dissatisfaction into enhanced perceptions, trust, and use intentions. Researchers and designers should adopt a longitudinal perspective spanning multiple encounters when evaluating and optimizing user experience.

7.5.2.6 Video Medium in eiIBM_RobotV2 Showing Greater Positive Interpretation Change

Comparative analyses revealed participants in the video-guided condition (using either a telepresence robot in **eiIBM_RobotV1** or virtual avatar in **eiIBM_RobotV2**) reported greater pre-post increases in positive interpretation endorsement after completing **eiIBM_RobotV2** versus **eiIBM_RobotV1**. This finding suggests that the improved avatar medium specifically helped boost positive interpretation tendencies initially, in contrast to the telepresence robot. Notably, the significance of positive interpretation change persisted even after removing outliers.

One potential explanation for this difference is the transfer of behavioral patterns. As the relationship between the user and the robot develops, the robot's behavioral patterns may be mirrored by the user, leading them to actively choose this mode of behavior, either verbal (Brandstetter et al., 2017) or non-verbal (Mutlu et al., 2009). This phenomenon is more likely to occur when users perceived the robot as more human-like and experience a sense of realism, as explained by the robot-mediated communication channel (Hoorn, 2020a). In the previous **eiIBM_RobotV1** study, many participants reported feeling detached from the role of the physical robot, likening the experience to acting in a play, However, **eiIBM_RobotV2** addressed this issue, fostering a more authentic interaction.

Interestingly, the SST measure did not exhibit the same replication effect. This may be attributed to the fact that SST also involved attention bias, which was not specifically targeted for change in the current exercise design.

These findings encourage researchers to consider the persuasive effect and transfer effect of mimicking the robot's behavioral pattern in CBT exercises incorporated by the robot. It would be a success of the far transfer effect of the **eiIBM** into the real-life, and thus reduce the depressive symptoms.

7.5.2.7 Tolerance Mitigating Mistake Effects

The findings provided preliminary evidence that higher initial tolerance mitigates the negative impact of perceiving more frequent robot mistakes on user satisfaction. Participants with greater tolerance may maintain higher satisfaction despite noticing more mistakes from the robot. Tolerance did not significantly correlate with the perceived frequency of mistakes, suggesting tolerant participants simply have fewer negative reactions when errors occur.

One explanation is that tolerant individuals exhibit less intense negative emotional responses like frustration or annoyance to the robot's mistakes compared to non-tolerant participants. Providing justifications for errors during the interaction may have pre-emptively reduced frustration for more tolerant users. Their tolerance enabled them to appraise mistakes more rationally, as correctable learning opportunities rather than catastrophes (Mirnig et al., 2017).

In contrast, less tolerant participants may react more negatively when perceiving errors, diminishing their satisfaction. They may view mistakes as reflecting permanent, global incompetence rather than situational, specific learning needs. These maladaptive appraisals elicit stronger negative emotions that erode user experience.

These findings indicate that higher dispositional tolerance shields user satisfaction from the detrimental impacts of perceiving errors during human-robot interaction. Fostering realistic expectations regarding occasional mistakes may further mitigate dissatisfaction. Optimizing tolerance and managing expectations are important considerations when incorporating imperfect AI systems into mental healthcare applications.

7.5.2.8 Trust Base Dimensions

Trust formation involves both the trustor's psychological tendencies and the trustee's perceived qualities (Mayer et al., 2006; Hancock et al., 2023). Participants with more severe

depression may have been more influenced by their own subjective characteristics in developing trust towards the robots. Their increase negative affect and reduced reward responsiveness could have led them to rely more on the emotional experience during the interaction when judging trustworthiness. In contrast, those with less severe depression were potentially less affected by their internal states and could leverage factual knowledge about the robot's capabilities alongside the emotional experience when determining trust. These findings highlight the shaping of trust in robots based on trustor characteristics, with more depressed individuals relying relatively more on emotional aspects versus factual knowledge.

7.5.3 Limitation and Future Works

While providing valuable insights, this study had some limitations that should be addressed in future research. In the comparative analyses between Study 1 and the present study, Study 1 was conducted online whereas the present study involved in-lab experiments. The platform and environment could influence user experience and outcomes. Future work should evaluate the **eiIBM_RobotV2** program in in-lab settings to examine its effects on experience and therapy outcomes in more naturalistic contexts.

While the current studies employed close supervision and restricted communication to ensure safety and confidentiality, the long-term scalability and at-home usage of such robot-delivered interventions necessitates more robust data management solutions. In the future, the author plan to leverage her lab's ongoing development of an offline storage system called the Robot Brain Server (RBS) and develop offline large language model. This custom-built platform will provide the author with greater control over participant data, allowing to securely store and process information without relying on external services.

Potential floor/ceiling effects caused from moderacy bias or central tendency bias (Saris & Gallhofer, 2007) on measurement scales may have made it difficult to detect subtle

variances in experience effects on outcomes. Without direct exposure to **eiIBM_RobotV2**, Study 1 participants' ratings may have clustered towards the scale midpoint as they lacked a reference point for comparison. The difference target constructs between studies, such as program use intention (*UseIntP*) versus robot use intention (*UseIntR*), could also limit comparability, Caution is needed when interpreting similarities and differences.

Unavoidable errors in experimental procedures, such as delayed response times from researcher-requested regeneration, could shape user perceptions. Future implementations should optimize system response latency and transparency to minimize perceptions of errors.

Chapter 8: General Discussion and Conclusion

This thesis serves as a starting point for scientifically understanding how to integrate social robots into psychotherapy, beginning with the elaborative interpretation bias modification (**eIBM**), which benefits depressed individuals. This chapter summaries the findings to the research questions which has been elaborated in Chapters 5 - 7 and discusses implications for integrating robots into the **eiIBM** program based on the three studies' findings. The chapter also discusses potential implications for online robot-delivered therapy design for the Hong Kong depressed young adults based on the major findings. Lastly, this chapter explains the limitations of the present study and suggests potential future research directions.

8.1 Summary of Key Findings

The elaborative interpretation bias modification (**eiIBM**) exercise format and the robot modalities used to deliver the intervention, along with their behaviors, play a crucial role in shaping user experience and cognitive effects related to depression relief. The iterative design of the **eiIBM** program incorporates imagery-enhanced elaborative interpretation bias modification, providing a structured and engaging framework for users to counter their negative interpretation tendencies. The enhanced imagery component and robot guidance contribute to improvements in interpretation bias and depressive symptoms reduction, as evidenced by the findings across the studies 5, 6 and 7.

Robot modalities, including text, audio, and video, offer varying levels of social presence cues and interactivity, influencing users' perceptions and experiences of the **eiIBM** program. These perceptions and experiences are derived from two proposed goal-relevant affordances: intervention delivery and ease-of-use interaction. The text modality (i.e., text-based chatbot) provides a straightforward and accessible means of delivering intervention

content, albeit with limited social cues. The audio modality introduces a human-like voice, reducing reading workload and mind wandering, leading to increased concentration and accompanying benefits. The video modality adds a visual representation of the robot and its behaviors, offering companionship and the highest level of interactivity.

The valence of affordances, compared to their relevance, contributes to the intention to use the program, with the valence of emotional distress relief playing a dominant role. This results from different aspect-focused evaluations derived from the three modalities, suggesting the need for improvements in the exercise format and robot behavior to align with the relational agent role in the therapy context. Consequently, the traditional version of the standardized **eiIBM** program was optimized into a variant by adding an automatic response system and virtual avatar, leading to improved user experiences and cognitive outcomes in the video modalities between the traditional version and the variant.

Although no single modality demonstrated superiority in user experience and intervention outcomes, the audio and video modalities in the robot benefit long-term use intention. As users become more experienced with the visual modalities, their initial oversimplifications and skepticism diminish, leading to greater acceptance. Compared to the audio robot, the emotionally aroused telepresence robot with visual modality makes users more likely to develop an affective relationship with the robot, as evidenced by the significant moderation effect of the medium on the affordance contributing to the use intention of the robot through valence (see Chapter 6). Moreover, with the prominent relational role of the robot in the **eiIBM** exercise, the valence of the robot overwhelms the relevance in explaining the use intention of the robot (see Chapter 7).

8.2 Methodological and Theoretical Contributions

This research makes several methodological and theoretical contributions to the field of human-robot interaction (HRI) and the application of social robots in mental health interventions. Methodologically, this study starts from the application of the Interactively Perceiving and Experiencing Fictional Characters (I-PEFiC; Hoorn, 2018) model to guide assessments of user experiences with robot-delivered mental health interventions. The popular TAM and UTAUT series models are not applicable for the relational agents in the current study, as those models' predictors were irrelevant for affective decisions (Gursoy et al., 2019) or goal-motivated cognitive decisions (Bagozzi, 2007). By operationalizing key constructs from the I-PEFiC framework, such as task (i.e., emotional distress relief) and interaction (i.e., ease-of-use) affordances, use intentions, and trust, this research demonstrates how the model can be leveraged to systematically evaluate user perceptions and predict intervention outcomes in the context of robot-assisted therapy.

The I-PEFiC model posits that people assess robots for their action possibilities (affordances) and compare them to personal goals and concerns (relevance and valence). Therefore, it is crucial to consider the specific context and characteristics of the robot when evaluating user experiences, as different robots may elicit distinct sets of expectations and evaluative criteria. The longitudinal design, involving repeated measurements of experiential variables across multiple sessions, aligns with I-PEFiC's emphasis on the dynamic nature of human-robot relationships. This methodological approach captures the evolution of user perceptions over time, providing a more comprehensive understanding of the interaction effects.

A multi-method approach, combining quantitative and qualitative data, was employed to provide a nuanced examination of the relationships between user perceptions, robot characteristics, and therapeutic outcomes. Various measures were utilized, including experiential variables, depressive symptom scales (e.g., BDI-II), and interpretation bias

screening tasks (e.g., WSAP-D, SRT, SST), to assess the impact of robot-delivered interventions. The iterative design of the study, starting from the framework and theory to test the simple incorporation of the robot, provides evidence that the simple incorporation of the robot did not harm the effectiveness of the **eiIBM** intervention. This step-by-step approach facilitates the understanding of the robot's role and guides the improvement of robot characteristics and relational role-fit in the **eiIBM** exercise, leading to the design optimization of the online robot-delivered therapy.

Theoretically, the findings validate and extend key propositions of the I-PEFiC model in the domain of robot-delivered mental health interventions. The strong influence of intervention delivery affordances on use intentions and trust, particularly among participants with higher levels of emotional distress, underscores the importance of designing robots that effectively address primary user goals. This task-contingency effect, where the robot's ability to meet the user's core needs shapes overall perceptions and engagement, aligns with I-PEFiC's emphasis on the role of affordances in driving user responses (van Vugt et al., 2009; Hoorn, 2015b). Furthermore, the observation that experiential evaluations evolve over time, as users gain more understanding of the robot's capabilities, supports the I-PEFiC's proposition that perceptions are recursively updated based on ongoing interactions (van Vugt et al., 2009; Hoorn, 2015b).

Moreover, the association between positive user experiences and reduced negative interpretation biases and depressive symptoms validates the assumption that user experience plays a crucial role in shaping intervention outcomes. The cluster analysis, showing the greatest improvements among participants with the most positive experiences, highlights the importance of fostering favorable perceptions to optimize the therapeutic effects of robot-delivered interventions. However, it is important to note that this finding is based on the rational agent-like behavior of the rigid robots in Study 1 and Study 2. One potential

explanation for the disappeared effect of experience on intervention outcomes is that the opened-ended exercise design ensures user participation and emotional engagement, leading to active training in either cognitive or affective processing which is critical for cognitive bias modification (Hoppitt et al., 2010). However, there remains another possibility that the natural verbal conversation with the relational agents engages participants self-disclosure even if they evaluate the experience as medium, as indicated by Luo et al. (2022).

8.3 Characteristics of Depressed Young Adults Appeared from eiIBM Exercise.

The **eiIBM** exercise, guided by robots, revealed several key characteristics of depressed young adults. These findings highlight the importance of personalized, adaptable, and emotionally sensitive approaches when integrating robots into mental health interventions.

8.3.1 Diverse Needs and Expectations of Robot Roles

Depressed young adults exhibited varying needs and preferences during the **eiIBM** exercise. Some participants sought more interactive and personalized experiences, while others preferred a straightforward and structured approach. Participants also held different expectations of the robot's role, with some viewing it as a helpful guide or companion and others perceiving it as a tool or extension of the intervention.

For participants with tangible triggers, a structured approach featuring clear, simple interactions and fewer modalities may be most beneficial. In contrast, those with intangible triggers may require a focus on companionship and a sense of interaction, rather than straightforward exercises. Clearly defining and communicating the robot's intended role and capability is crucial to align with user expectations and avoid the expectancy violation effect (Go & Sundar, 2019), which might lead to positive valence.

8.3.2 Tolerance for Robot Errors

Depressed young adults demonstrated some tolerance for robot errors during the **eiIBM** exercise. While technical glitches and inconsistencies in the robot's responses could cause frustration, participants generally remained engaged. Self-admitted and self-corrected errors by the robot can help reduce the negative impact (Klüber and Onnasch 2022), as this active vulnerability may elicit empathy from users. According to Maslow's hierarchy of needs (Maslow,1954), active vulnerability can make users feel valued, and when this psychological need is met, they are more likely to understand the robot's difficulties from its perspective. Aligning expectations about the robot's error-making capabilities with its actual performance could avoid the negative affect due to misplaced trust or frustration (Schramm et al., 2020), particularly for less tolerant participants.

8.3.3 Emotional Sensitivity and Vulnerability

The studies revealed that participants' trust criteria in the robot guide varied based on their depression severity. Less depressed individuals tended to build trust based on previous experiences, while severely depressed individuals relied more on current experiences. This finding underscores the emotional sensitivity and vulnerability demonstrated by depressed young adults during the **eiIBM** exercise, for instance, negative rumination to the negative stimuli (Nolen-Hoeksema, 1991; Kellough et al., 2008). Careful design and implementation of robot-assisted interventions are essential to minimize potential harm and provide appropriate emotional compensation for this vulnerable population.

In summary, the characteristics of depressed young adults in the **eiIBM** exercise emphasize the heterogeneity among this population and the need for personalized, adaptable, and emotionally sensitive approaches when integrating robots into mental health interventions.

8.4 Research implications

The findings from this thesis have several important implications for research and practice in the field of human-robot interaction and digital mental health interventions.

8.4.1 Implications for eiIBM Exercise for Depressed Young Adults

In the **eiIBM** exercise examined in this thesis, the relational robots equipped with natural language processing capabilities were able to demonstrate to participants how to interpret ambiguous situations in a more positive manner. The variant of **eiIBM** exercise (i.e., **eiIBM_Robotv2**) with open-ended resolution fill-in component seems to be independent from the experience effect, due to its compulsory to actively generate resolution. Specifically, the avatar guide in **eiIBM_RobotV2** led participants through the process of positive interpretation, serving as a social model that participants may have emulated in their own behaviors (Brandstetter et al., 2017; Mutlu et al., 2009). This helps to explain why the avatar-guided **eiIBM_RobotV2** elicited more elaborated positive interpretation endorsements from participants compared to the telepresence robot in **eiIBM_RobotV1**, which simply asked participants to provide resolutions without providing a demonstration.

Robots guiding **eiIBM** exercise evoke strong affective responses in humans, making them more likely to follow the robot's guidance compared to rational media (Lopez et al., 2017; Ghazali et al., 2019), such as text-based mental health interventions. By leveraging the modalities and interactivity of a robot guide, **eiIBM** exercises can become more engaging and motivating for depressed young adults.

8.4.2 Relationships Between eiIBM Exercise Format and Robot Modalities with User Behaviors

The findings from this thesis illuminate the intricate relationships between the **eiIBM** exercise format, robot modalities, and user behaviors in the context of online mental health

interventions for depressed young adults. The results demonstrate that while the **eiIBM** exercise itself is effective, the introduction of a relational agent, such as a robot, can significantly influence users' expectations and perceptions of the intervention.

Firstly, the studies show that the presence of a robot as a relational agent alters users' social expectations and their evaluation of the **eiIBM** exercise format. In the text condition, users perceived the guide as an automated program and did not form strong opinions or expectations. However, as the robot's modality progressed to audio and video conditions, users' evaluations became more complex, and their satisfaction with the existing **eiIBM** format decreased [Chapter 7, Section 7.4.2]. This finding suggests that the incorporation of a more socially present robot can lead to higher expectations and a desire for a more engaging and personalized intervention experience (Sandoval et al., 2014).

Secondly, the results indicate that a more realistic or human-like social robot does not necessarily lead to better outcomes. When the robot is perceived as a relational agent and the **eiIBM** exercise as a method, the relationship between the two can be compared to that of a therapist and a therapy. In such cases, users' hesitation or enthusiasm towards the therapy may not immediately affect their short-term liking of the therapist or their willingness to continue the intervention. However, if users dislike the therapist, it can lead to disengagement from the therapy [Chapter 6, Section 6.4.3]. Furthermore, the findings suggest that when the valence towards the robot is too high, users may perceive the robot as having a human presence behind it, potentially triggering the human-human social rewards system (Zhang et al., 2021). This can result in difficulty trusting the encouragement and affirmation provided by the robot, as it may be seen as coming from a human rather than an objective, automated system [Chapter 7, Section 7.4.3].

These findings have important implications for the design of robot-delivered mental health interventions. While the incorporation of relational agents can enhance user

engagement and satisfaction, designers must carefully consider the balance between the robot's social presence and its perceived role in the intervention. A highly realistic or human-like robot may not always be the most effective choice, as it can lead to unmet expectations and a breakdown in trust if users perceive the robot as a human-controlled entity (Sandoval et al., 2014).

8.4.3 Implications for Developing Online Robot-Delivered Therapy in Hong Kong

The findings from this thesis have important implications for the development and implementation of online robot-delivered therapy in Hong Kong, particularly for depressed young adults. The results indicate that **eiIBM_RobotV2**, with its improved design based on Study 2 and enhanced interactive communication capabilities, has greater long-term benefits than **eiIBM_RobotV1**. This superiority can be attributed to three key factors: (1) the incorporation of user feedback and expectations from Study 2, (2) the natural language processing capabilities of the robot in **eiIBM_RobotV2** that enable free conversation, and (3) the more humanoid robot's ability to elicit realistic interactions and encourage mimicry behavior (Hoorn, 2020a; Brandstetter et al., 2017; Mutlu et al., 2009).

However, it is essential to consider the potential negative impact of a robot being too human-like. Providing users with a choice of robot modalities is crucial, as different individuals may have varying preferences for experiencing realism in the therapeutic context (Scassellati et al., 2018). No single robot modality is inherently superior; offering options allows users to select the modality that best suits their needs without compromising the intervention's efficacy.

Developing an online version of **eiIBM_RobotV2** could have significant social impact by increasing access to mental health support for depressed young adults in Hong Kong. The city faces a shortage of mental health professionals, with only 4.4 psychiatrists per 100,000

people, compared to the global median of 8.6 (WHO, 2023). This scarcity, combined with the stigma surrounding mental health, creates barriers to accessing timely support. By integrating a social agent like a robot into their daily routines, depressed young adults can receive timely support in reframing their negative interpretations of real-life situations, ultimately helping them to adjust their negative patterns.

However, implementing online robot-delivered therapy in Hong Kong requires careful consideration of cultural, ethical, and occupational issues. Ensuring user privacy is a significant challenge, as depressed individuals may be particularly concerned about the confidentiality of their interactions with the robot. At the same time, an alert system should be in place to allow social workers and counsellors to intervene when users display suicidal tendencies.

Another challenge is the potential for robot errors to negatively impact users, especially when scaling up the intervention without sufficient human oversight. One approach to mitigate this risk is to engage volunteers in providing simple feedback and guidance, reducing training costs and enabling broader participation. The crowdsourced feedback could be used to supervise and strengthen the robot's model, adapting it to the Hong Kong societal needs. However, this approach also raises privacy concerns that must be addressed.

It is important to recognize that **eiIBM** may not be suitable for all depressed individuals, as some may not exhibit negative interpretation biases but other cognitive bias (e.g., attentional bias and memory bias). To cater to diverse needs, the robot should be designed to deliver a variety of therapies. The work presented in this thesis serves as a starting point, highlighting the need for future studies to explore the delivery of personalized, multi-modal therapies via social robots.

Different social groups may have varying attitudes towards the use of robots in mental health care, influencing their acceptance and adoption of the technology through the evaluation on ethics (good or bad) (Van Vugt, Konijn, & Hoorn, 2009). Researchers should engage stakeholders, such as mental health professionals, policymakers, and potential users, to ensure that the robot-delivered therapy aligns with local needs, values, and regulations. Additionally, efforts should be made to educate the public about the capabilities and limitations of social robots in therapy, dispelling oversimplifications and fears of robots replacing human therapists.

Given Hong Kong's advanced information infrastructure, there is an opportunity to capitalize on the city's connectivity to deliver timely and personalized support to depressed young adults. By integrating robots into their daily digital routines, these individuals can receive ongoing assistance in reframing their negative interpretations of real-life situations, ultimately helping them to adjust their negative patterns.

8.4.4 Generalizability of the Findings

The findings from this research have significant potential for broader applicability across various robotic systems and mental health interventions. While the specific voice and appearance characteristics were investigated, the core principles identified can be generalized to a wider range of contexts.

A key observation from this study was the prominence of the emotional distress relief affordance over the ease-of-use affordance for the depressed young adult participants. They seemed to prioritize the robot's ability to provide emotional distress relief over its ease of use. This finding aligns with research on social assistants providing target service/function (e.g., automatic driving assistant; Sun, 2023), where perceived ease of use had minimal impact on usage intention before the primary need was met. Sun (2023) suggests that the service robot's

ability to self-admit and self-correct errors is distinct to other types of self-service programs. Unlike self-service technologies with predefined functions, service robots can engage in dialogues with users to facilitate task completion. This dialogue-based approach allows users to provide feedback and work through any issues collaboratively, potentially mitigating the negative impact of errors (Sun, 2023). As long as the robot maintains its relational role and uses appropriate strategies to address errors, such as self-admission and self-correction, the ease-of-use aspect may become less critical in the long run. These insights suggest the potential for generalization to diverse robotic platforms, such as virtual robots and audio bots with different voices and appearances. However, it is important to note that the generalizability of the findings may be limited when considering physical robots in therapy, as some studies have suggested that depressed individuals may not prefer embodied robots, preferring virtual or text-based interactions instead (Scassellati et al., 2018).

Also, it does not mean the ease-of-use affordance should be disregarded, as it can lead to user early drop-out if the young adults did not find the interaction pleasant. Additionally, if the modality characteristics affect the task completion perception, it would also harm the acceptance. Therefore, if a robot can effectively deliver the therapy content and benefit the young adults, the specific modalities used may not be a significant barrier to acceptance, as long as they align with the robot's intended role and facilitate task completion, making the generalization among the other robots more supportive.

Regarding the generalization of the therapy method, it remains uncertain whether the principles would apply to robot-delivered unstandardized therapies, such as counseling. Role theory (Solomon et al., 1985) suggests that when both the user and the robot act according to socially defined roles, role consistency is more likely to occur (Sun, 2023). Conversely, if the robot does not conform to its prescribed role, role inconsistency may arise and negatively affect the user's perceptions. This implies that when the robot is expected to take on a more

clinical role, such as that of a therapist, the requirements for empathy, understanding, and professional expertise may be higher than what current social robots can reliably provide. In these cases, more research is needed to explore the appropriate boundaries and integration of robot-assisted therapy, where the robot may serve more as a complementary tool rather than a primary therapeutic agent. But for the standardized therapies, the role consistency in the current study might support the findings generalization.

Overall, the key principles and insights from this research have significant potential to inform the design and implementation of robot-assisted mental health interventions, particularly for standardized therapies, while the generalization to unstandardized contexts requires further investigation. The emphasis on emotional distress relief and role alignment are crucial considerations for effective robot-assisted therapy.

8.5 Limitations and Future Research

While the limitations specific to each study have been discussed in their respective chapters, this thesis has several overarching limitations due to social issues, limited time, and research scope.

First, the scales used in Study 1 were designed based on assumptions about users' goals, such as emotional relief and ease of use in interaction. However, as discovered in Study 2, the content of ease-of-use intention may vary for individuals with diverse needs. If the order of Study 1 and Study 2 were reversed, the scales used in Study 1 could have been better designed to capture these nuances. To ensure comparability of user experiences between Study 1 and Study 3, the scales mainly remained unchanged in Study 3. Future experiments should further specify the content of valence and relevance based on the findings from Study 2, as different user groups may create different social constructs around robots, influencing their discrepancy between perceived affordances and designed affordance.

Additionally, future studies could consider administering the full WSAP to participants after the intervention, as the present study limited its item number to avoid participant fatigue.

Second, while age, gender, and depression level were controlled for in the experiments, the study did not fully account for potential confounding events that may have occurred during the experimental period. Given the relatively small sample size, these uncontrolled factors could have influenced the results. Further research should aim to replicate these studies with larger and more diverse samples to enhance the external validity of the findings. Additionally, the studies did not explore or control other potential cofounding variables, such as prior experience with (social) robot and mental health intervention, social ethnic group and individual differences in emotional regulation strategies. Accounting for these additional factors could provide a comprehensive understanding of the underlying mechanisms influencing relational agent acceptance.

Third, the study design did not include a placebo or a baseline comparison, such as a human-delivered intervention or other CBT therapy. This makes it challenging to fully attribute the observed experiences and outcomes solely to the different robot modalities. Further studies should consider incorporating such control conditions to better isolate the effects of robot-mediated interventions.

Fourth, verbal data collected at the end of the conversation with the robot in Study 3, where users expressed their usage experience and reasons for satisfaction, was not included in the analysis due to time constraints during the Ph.D. thesis stage. This data could have provided additional insights into the results. During the review stage, the researcher plans to analyze this data to gain a deeper understanding of user perceptions and experiences.

Final, although the effects of changes in the **eiIBM** exercise and the **eiIBM** robot were analyzed, confident comparisons between the results of Study 1 and Study 3 cannot be made

due to significant differences in the experimental environment and platform. Study 2 showed a significant but weak correlation between the use intentions of robot and of program.

Looking onward, the robot developed in this thesis was not yet highly intelligent. Further tests are planned to understand user interaction experiences and therapeutic effects in natural settings. Additionally, the researcher aims to incorporate robots into a wider range of therapies to explore their potential in supporting various mental health interventions.

Ultimately, the researcher's goal is to develop a fully autonomous language model intervention to assist depressed individuals whenever possible without professional engagement. This will expand access to mental health support and empower individuals in their journey towards well-being. However, achieving this goal requires addressing several research challenges, such as demarcating the degree to which a robot should simulate human behaviors, determining the robot's social position, and understanding how users' perceptions of robots evolve with experience and appropriation.

APPENDICES

Appendix A: Ethic Approval

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

To	Hoorn Johannes Ferdinand (School of Design)		
From	Ng Po Ling, Delegate, Departmental Research Committee		
Email	bobo-pl.ng@	Date	03-Aug-2022

Application for Ethical Review for Teaching/Research Involving Human Subjects

I write to inform you that approval has been given to your application for human subjects ethics review of the following project for a period from 01-Jul-2022 to 30-Jan-2024:

Project Title:	Research on Negative Bias Modification among Greater Bay Area youth: Design and Applications
Department:	School of Design
Principal Investigator:	Hoorn Johannes Ferdinand
Project Start Date:	01-Jul-2022
Project type:	Human subjects (non-clinical)
Reference Number:	HSEARS20220730001

You will be held responsible for the ethical approval granted for the project and the ethical conduct of the personnel involved in the project. In case the Co-PI, if any, has also obtained ethical approval for the project, the Co-PI will also assume the responsibility in respect of the ethical approval (in relation to the areas of expertise of respective Co-PI in accordance with the stipulations given by the approving authority).

You are responsible for informing the PolyU Institutional Review Board in advance of any changes in the proposal or procedures which may affect the validity of this ethical approval.

Ng Po Ling

Delegate

Departmental Research Committee (on behalf of PolyU Institutional Review Board)

Appendix B: Main Analyses and Their Power Analysis

Main Analyses in Study 1

Analysis 1: Understand the Effect of Robot Modalities on Experience

The study employed Generalized Estimating Equations (GEE) to assess the impact of Time (*T1* vs *T2* vs *T3*) and Medium (Audio vs Video vs Text) on participants' eight experiential variables: *M_AffEase*, *M_RelEase*, *M_ValEase*, *M_ValEmo*, *M_UseIntP, M_TrustP, M_UseP_i*, and *M_UseP_ci*.

The sample size and statistical power were calculated using the PASS software (Ahn,2015; Zhang & Ahn,2013). The analysis was based on a design with three groups, each containing approximately 16 clusters, with measurements taken at three time points per cluster. The group means, standard deviations (*SD*), and intracluster correlation coefficients (*ICC*) for each variable were calculated and are presented in Table A.1.

When the calculated values were input into the PASS software, the results revealed that the statistical power for all the variables was very low, ranging from 0.07 to 0.20. This low power is likely due to a combination of small sample size and uncertainty in parameter estimates. Specifically, the study did not have the benefit of prior literature or pilot data to reliably estimate the group means, standard deviations, and ICCs. These parameters were calculated based on the current data, which may have resulted in less accurate power estimates.

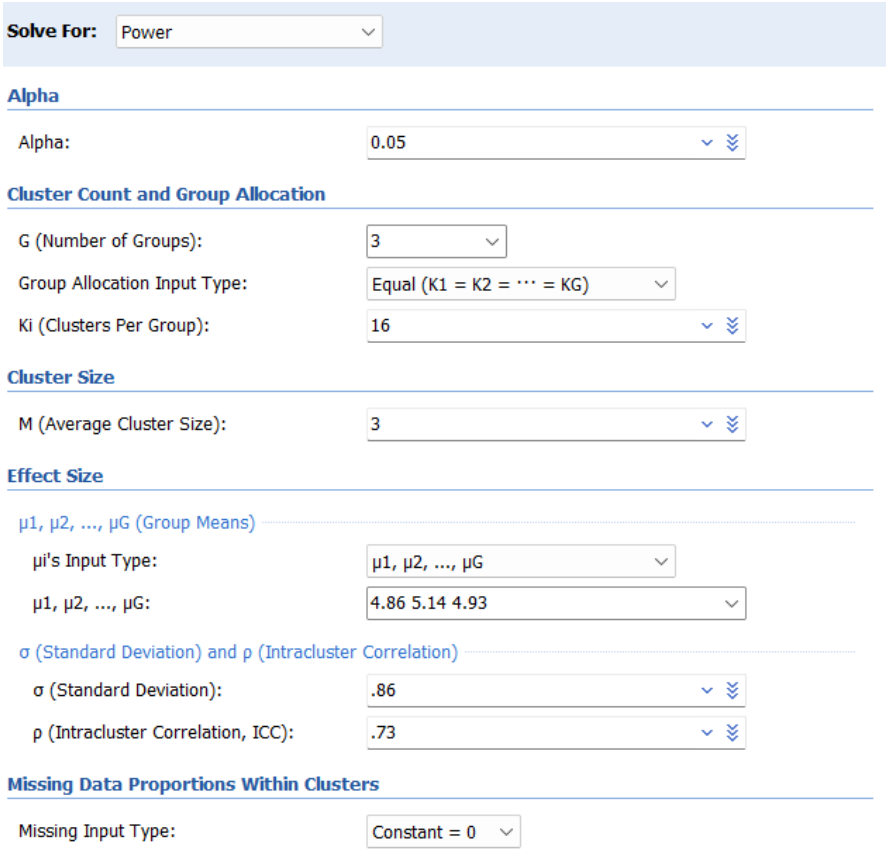

Figure A.1. Demo of PASS Calculating Power Evaluating Experiential Data in Study 1

Table A. 1 Parameters for PASS Calculating Power Evaluating Experiential Data in Study 1

Item	**μ1, μ2, ..., μG**	**σ (Standard Deviation)**	**ρ (Intracluster Correlation, ICC)**	**Power**
M_AffEase	4.86 5.14 4.93	.86	.73	.14
M_RelEase	4.53 4.71 4.66	.91	.67	.08
M_ValEase	4.45 4.65 4.62	.89	.75	.09
M_ValEmo	4.23 4.32 4.41	.91	.69	.08
M_UseIntP	4.08 4.29 4.44	.91	.64	.19
M_TrustP	4.27 4.65 4.38	.95	.68	.20
M_UseP_i	4.52 4.63 4.65	.93	.58	.07
M_UseP_ci	4.53 4.72 4.74	1.03	.65	.09

Note: μ1, μ2, ..., μG: group mean of the variables; σ (Standard Deviation): the standard deviation of the responses within a cluster; ρ (Intracluster Correlation, ICC): the intracluster correlation coefficient within groups

Analysis 2: Effect of Robot Modalities on Therapy Outcome

The author employed Generalized Estimating Equations (GEE) to examine the interaction effects between **Medium** (*Audio*, *Video*, *Text*, *Control*) and TestTime (*pretest*, *posttest*) on six assessment measures: *DS_C*, *SST_TNR*, *WASP_NER*, *WASP_PMR*, *SRT_PT*, and *SRT_NT*. The key parameters used for calculating the statistical power of the study, including the group means (μ1, μ2, ..., μG), standard deviations (σ), and intracluster correlation coefficients (ρ) for each assessment measure, are presented in Table A.2.

The calculated statistical power for the different measures ranged from a low of 0.11 for *SRT_NT* to a high of 0.67 for *WASP_NER*. Similar to the previous analysis, the relatively low statistical power is likely due to the small sample size, with approximately 16 clusters per group, as well as the uncertainty in the parameter estimates, as the study did not have the benefit of prior literature or pilot data to reliably estimate the group means, standard deviations, and ICCs.

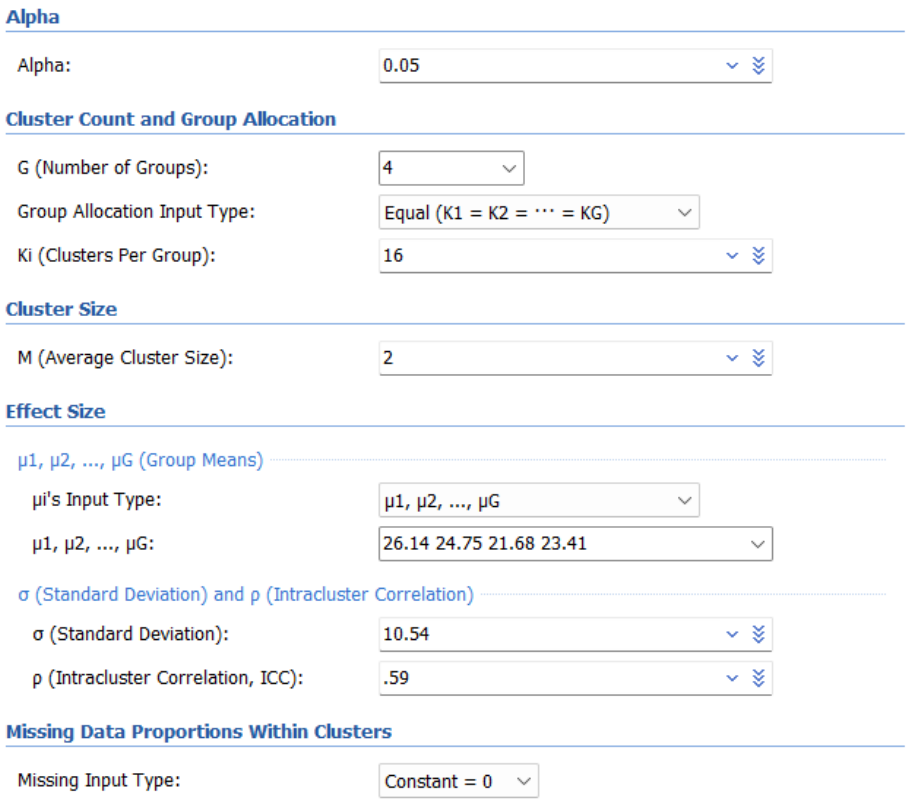

Figure A.2. Demo of PASS Calculating Power Evaluating Assessment Data in Study 1

Table A. 2 Parameters for PASS Calculating Power Evaluating Assessment Data in Study 1

Item	μ1, μ2, ..., μG	σ (Standard Deviation)	ρ (Intracluster Correlation, ICC)	Power
DS_C	26.14 24.75 21.68 23.41	10.54	.59	.19
SST_TNR	0.60 0.51 0.47 0.47	0.26	.54	.31
WASP_NER	0.65 0.56 0.49 0.50	0.22	.38	.67
WASP_PMR	0.48 0.36 0.39 0.38	0.18	.51	.48
SRT_PT	23.58 27.41 27.68 24.50	6.02	.29	.69
SRT_NT	20.56 19.13 19.12 18.75	6.83	.47	.11

Note: μ1, μ2, ..., μG: group mean of the variables; σ (Standard Deviation): the standard deviation of the responses within a cluster; ρ (Intracluster Correlation, ICC): the intracluster correlation coefficient within groups

Analysis 3: Effect of Robot Modalities on Residual Change Scores

One-way MANOVA analyses on the residual change scores of different intervention groups (**Medium**: *Audio*, *Video* and *Text*) were conducted. The residual change scores are *RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, *RES_NT* and *RES_PT*. The author calculated the

G*Power, results (Figure A.1) shown that given a conventional rejection area ($p < .05$), a power of .80, and sample size of $N = 49$, the effect sizes were expected to be around $\eta_p^2 = .32$, which is acceptable for a first experiment on any research topic.

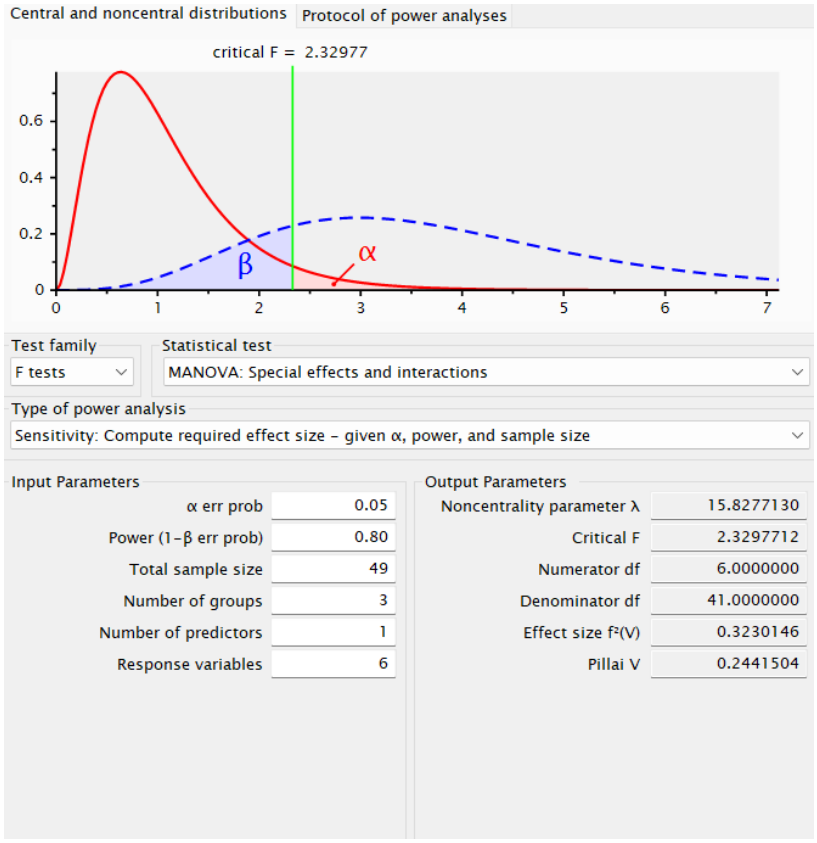

*Figure A.3. G*Power analyses, Computing Required Effect Size for Robot Modalities on Residual Change Scores (Study 1)*

Analysis 4: Effect of Experience on Intervention Outcome Difference

The experience clusters formed a new ranking variable *ExpRank* (*Exp_H*, *Exp_M* and *Exp_L*). A one-way MANOVA was conducted to assess the differences in the 6 intervention outcomes (*DS_MS*, *SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*) based on the ExpRank grouping.

Though the grouping variable was changed from *Medium* to *ExpRank*, the parameters filled in the G*Power analysis remained the same. Therefore, the effect size η_p^2 was again found to be .32, which is considered an acceptable effect size.

Main Analyses in Study 2

The study employed a within-subject design with **Medium** (*Video*, *Audio*, *Text*). Experience variables *M_AffEase*, *M_RelEase*, *M_ValEase*, *M_UseIntR* and *M_UseP* were measured after each Medium interaction.

As most data were non-normally distributed, nonparametric Friedman tests compared the three paired samples (*Text*, *Audio*, *Video*) on the experiential variables. However, since the Friedman Test is not implemented in G*Power, the authors used the parametric alternative and applied a correction to the results. Lehmann (2006) recommends that when using a nonparametric test, the researcher should first compute the sample size required for the parametric equivalent and then add 15% as an adjustment. Therefore, the authors reduced the calculated sample size by 15%, resulting in a final sample size of 30.

The authors then calculated the G*Power, and the results (Figure A.4) showed that given a conventional rejection area ($p < .05$), a power of .80, and a sample size of $N = 30$, the expected effect sizes were around $\eta_p^2 = .24$. This effect size is considered acceptable for a first experiment on the research topic.

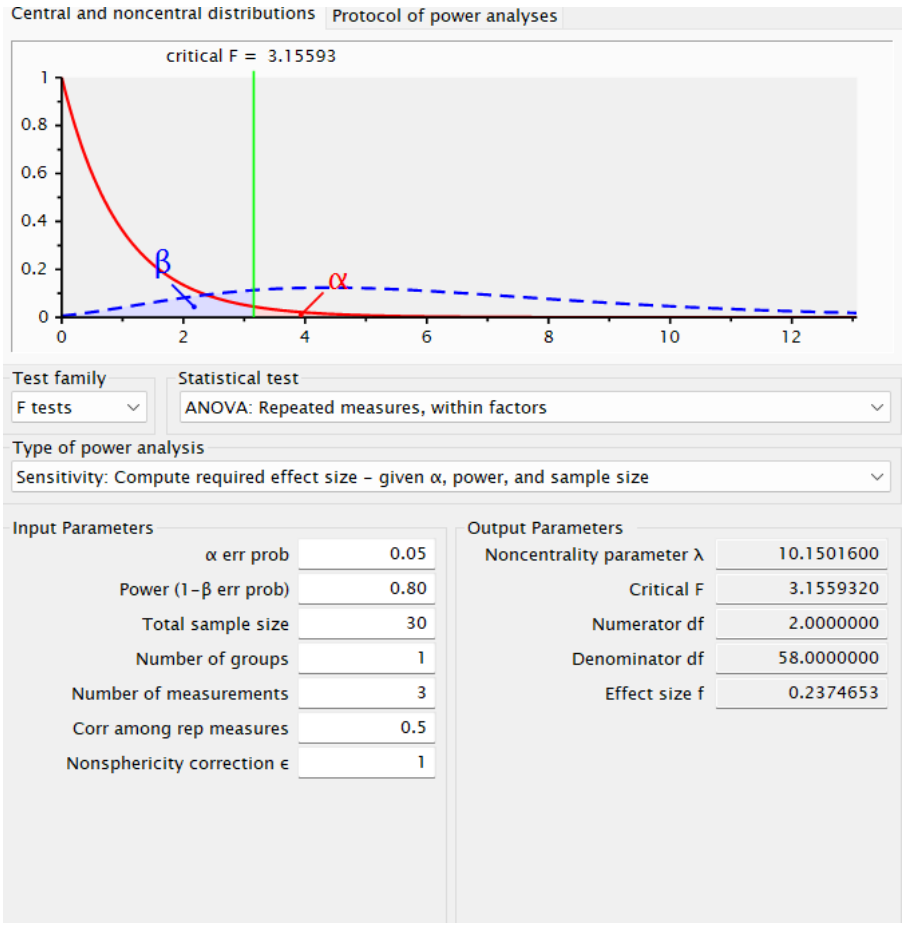

*Figure A.4. G*Power analyses, Computing Required Effect Size for Repeated Exposure (Study 2)*

Main Analyses in Study 3

Analysis 1: Understand the Effect of Robot Modalities on Experience

Generalized Estimating Equations (GEE) were employed to assess the impact of **Time** (*T1* versus *T2* versus *T3*) and **Medium** (*Audio* versus *Video*) on participant 6 experiences: *M_AffEase, M_RelEase, M_ValEase, M_UseIntR, M_TrustP, M_UseP*

Analysis 2: Effect of Robot Modalities on Therapy Outcome

For intervention outcome data, the focus of GEE was on interaction effects between **Medium (***Audio*, *Video*, *Control***)** and **TestTime (***pretest*, *posttest***)** in GEE models on six assessment measures: *DS_MS*, *SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*.

Analysis 3: Effect of Robot Modalities on Residual Change Scores

The residual change scores obtained were *RES_DS*, *RES_TNR*, *RES_NER*, *RES_PMR*, *RES_NT*, and *RES_PT*. One-way MANOVA analyses were conducted on the residual change scores across the different intervention groups (**Medium**: *Audio*, *Video*).

Using the same parameters as in Study 1 but with fewer groups, the G*Power analysis showed that the expected effect size for this study is $\eta_p^2 = .29$ (in Figure A.5). This effect size is considered acceptable for a first experiment on this research topic.

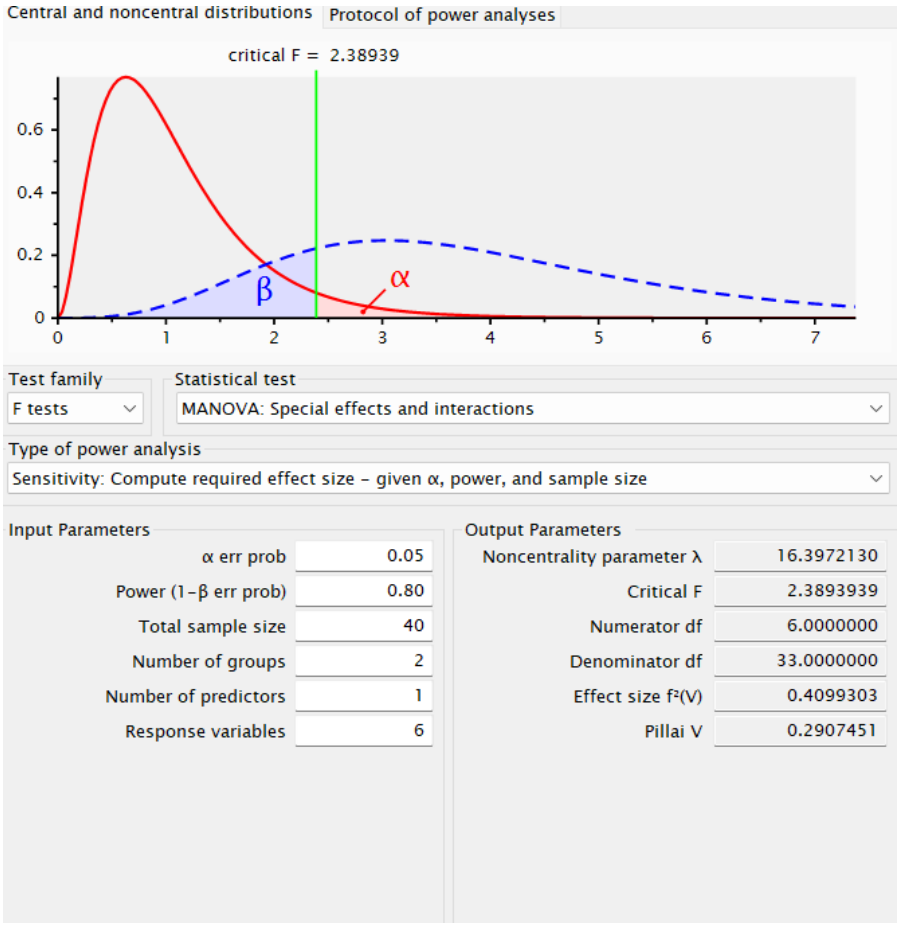

*Figure A.5. G*Power analyses, Computing Required Effect Size for Robot Modalities on Residual Change Scores (Study 3)*

Analysis 4: Effect of Experience on Intervention Outcome Difference

The experience clusters formed a new ranking variable *ExpRank* with three level: *Exp_H, Exp_M and Exp_L*. A one-way MANOVA was conducted to assess the *ExpRank* differences in six intervention outcomes (*DS_MS*, *SST_TNR*, *SRT_PT*, *SRT_NT*, *WSAP_NER* and *WSAP_PMR*) across the *ExpRank* groups.

The G*Power analysis, as shown in Figure A.6, indicated that with a conventional rejection area ($p < .05$), a power of .80, and a sample size of $N = 40$, the expected effect sizes were around $\eta_p^2 = .41$. This effect size is considered adequate for a first experiment on this research topic.

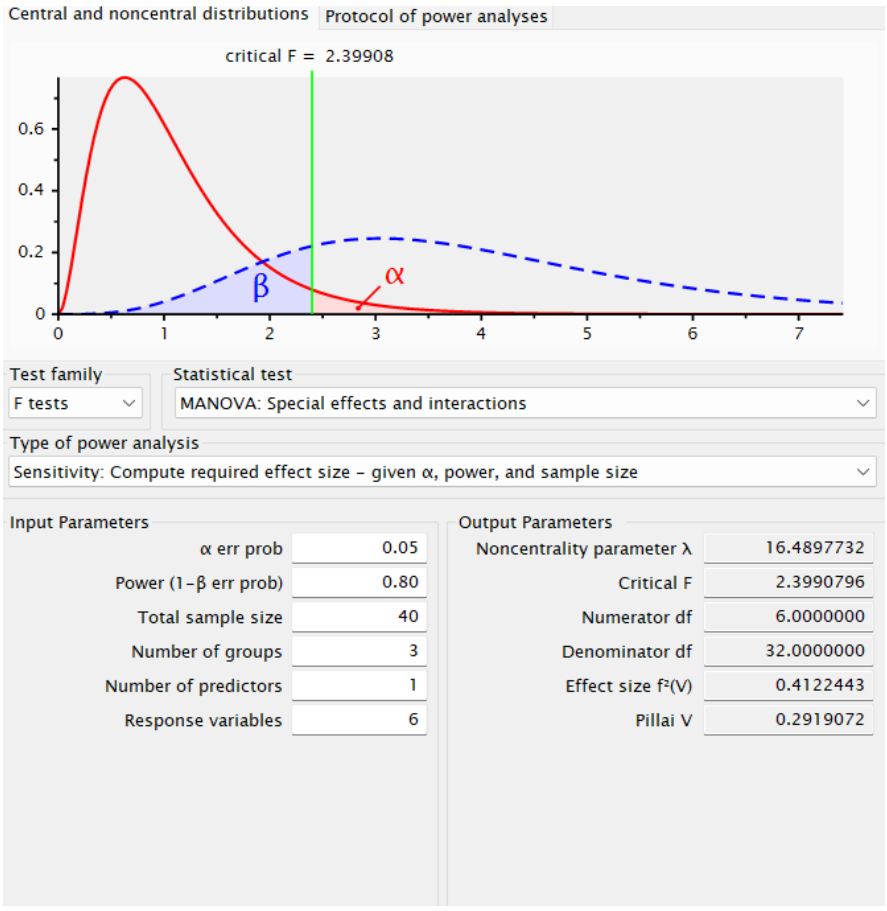

*Figure A.6. G*Power analyses, Computing Required Effect Size for Experience Rank on Residual Change Scores (Study 3)*

REFERENCE

Aguirre Velasco, A., Cruz, I. S., Billings, J., Jimenez, M., & Rowe, S. (2020). What are the barriers, facilitators and interventions targeting help-seeking behaviors for common mental health problems in adolescents? A systematic review. *BMC Psychiatry*, *20*(1). https://doi.org/10.1186/s12888-020-02659-0

Ahn, C., Heo, M., and Zhang, S. 2015. Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research. *CRC Press*. New York. See pages 116-119.

Ajzen, I. (1980). Understanding attitudes and predicting social behavior. *Englewood cliffs.*

Alemi, M., Ghanbarzadeh, A., Meghdari, A., & Moghadam, L. J. (2015). Clinical application of a humanoid robot in pediatric cancer interventions. *International Journal of Social Robotics*, *8*(5), 743–759. https://doi.org/10.1007/s12369-015-0294-y

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5). https://doi.org/10.1002/widm.1424

Amani, A., Shakarami, Z., Khezri, H. & Hadi, M. (2014). Investigation on Effects of Self-Esteem on High School Students' Academic Achievement. *International Journal of Basic Sciences & Applied Research*. 3. 312-315.

Anderson, J. R. (1976). *Language, memory, and thought*. Lawrence Erlbaum.

Arksey, H., & Knight, P. (1999). *Interviewing for social scientists: An introductory resource with examples*. Sage Publications.

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, *55*(5), 469–480. https://doi.org/10.1037//0003-066x.55.5.469

Arnett, J. J. (2024). *Emerging adulthood: The winding road from the late teens through the twenties*. Oxford University Press.

Bagozzi, R. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244–254. https://doi.org/10.17705/1jais.00122

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2010). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41–52. https://doi.org/10.1007/s12369-010-0082-7

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, 7(2), 127–150. https://doi.org/10.1177/1094428104263672

Bayraktaroglu, S., Kahya, V., Atay, E., & Ilhan, H. (2019). Application of expanded technology acceptance model for enhancing the HRIS usage in smes. *International Journal of Applied Management and Technology*, *18*(1). https://doi.org/10.5590/ijamt.2019.18.1.04

Beard, C., & Amir, N. (2009). Interpretation in social anxiety: When meaning precedes ambiguity. *Cognitive Therapy and Research*, *33*(4), 406–415. https://doi.org/10.1007/s10608-009-9235-0

Beard, C., Weisberg, R. B., & Primack, J. (2011). Socially anxious primary care patients' attitudes toward cognitive bias modification (CBM): A qualitative study. *Behavioral and Cognitive Psychotherapy*, *40*(5), 618–633. https://doi.org/10.1017/s1352465811000671

Beck, A. T. (1987). *Cognitive models of depression*. Journal of Cognitive Psychotherapy, 1(1), 5-37.

Beck, A. T. (1991). *Cognitive therapy and the emotional disorders*. Penguin.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, *165*(8), 969–977. https://doi.org/10.1176/appi.ajp.2008.08050721

Beck, A. T., & Bredemeier, K. (2016). A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clinical Psychological Science*, *4*(4), 596–619. https://doi.org/10.1177/2167702616628523

Beck, A. T., & Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual Review of Clinical Psychology*, *10*(1), 1–24. https://doi.org/10.1146/annurev-clinpsy-032813-153734

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory–II. *PsycTESTS Dataset*. https://doi.org/10.1037/t00742-000

Beck, A.T. (1963). Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, *9*(4), 324–333. https://doi.org/10.1001/archpsyc.1963.01720160014002

Beck, A.T. (1967). Depression. Harper and Row: New York.

Beck, J. S. (2005). Cognitive therapy for Depression. *PsycTHERAPY Dataset*. https://doi.org/10.1037/v00472-001

Beevers, C. (2005). Cognitive vulnerability to depression: A dual process model. *Clinical Psychology Review*, *25*(7), 975–1002. https://doi.org/10.1016/j.cpr.2005.03.003

Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2019). The next generation: Chatbots in clinical psychology and psychotherapy to Foster Mental Health – a scoping review. *Verhaltenstherapie*, *32*(Suppl. 1), 64–76. https://doi.org/10.1159/000501812

Berna, C., Lang, T. J., Goodwin, G. M., & Holmes, E. A. (2011). Developing a measure of interpretation bias for depressed mood: An ambiguous scenarios test. *Personality and Individual Differences*, *51*(3), 349–354. https://doi.org/10.1016/j.paid.2011.04.005

Bhattacherjee, A. (2001). Understanding Information Systems Continuance: An expectation-confirmation model. *MIS Quarterly*, 25(3), 351. https://doi.org/10.2307/3250921

Bibi, A., Margraf, J., & Blackwell, S. E. (2020). Positive imagery cognitive bias modification for symptoms of depression among university students in Pakistan: A pilot study. *Journal of Experimental Psychopathology*, *11*(2), 204380872091803. https://doi.org/10.1177/2043808720918030

Bickmore, T. W., Trinh, H., Olafsson, S., O'Leary, T. K., Asadi, R., Rickles, N. M., & Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google assistant. *Journal of Medical Internet Research*, *20*(9). https://doi.org/10.2196/11510

Bishop, S. J., & Gagne, C. (2018). Anxiety, depression, and decision making: A computational perspective. *Annual Review of Neuroscience*, *41*(1), 371–388. https://doi.org/10.1146/annurev-neuro-080317-062007

Blackwell, S. E., Browning, M., Mathews, A., Pictet, A., Welch, J., Davies, J., Watson, P., Geddes, J. R., & Holmes, E. A. (2015). Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults. *Clinical Psychological Science*, *3*(1), 91–111. https://doi.org/10.1177/2167702614560746

Blackwell, S. E., Schönbrodt, F. D., Woud, M. L., Wannemüller, A., Bektas, B., Braun Rodrigues, M., Hirdes, J., Stumpp, M., & Margraf, J. (2022). Demonstration of a 'leapfrog' randomized controlled trial as a method to accelerate the development and optimization of psychological interventions. *Psychological Medicine*, *53*(13), 6113–6123. https://doi.org/10.1017/s0033291722003294

Bower, G. H. (1981). Mood and memory. *American Psychologist*, 36(2), 129–148. https://doi.org/10.1037/0003-066X.36.2.129

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Bradley, J. V. (1958). Complete counterbalancing of immediate sequential effects in a Latin square design. *Journal of the American Statistical Association*, *53*(282), 525. https://doi.org/10.2307/2281872

Brandstetter, J., Beckner, C., Sandoval, E. B., & Bartneck, C. (2017). Persistent lexical entrainment in HRI. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2909824.3020257

Brettschneider, M., Neumann, P., Berger, T., Renneberg, B., & Boettcher, J. (2015). Internet-based interpretation bias modification for social anxiety: A pilot study. *Journal of Behavior Therapy and Experimental Psychiatry*, *49*, 21–29. https://doi.org/10.1016/j.jbtep.2015.04.008

Byrne, B. M., Stewart, S. M., & Lee, P. W. (2004). Validating the beck depression inventory-II for hong kong community adolescents. *International Journal of Testing*, *4*(3), 199–216. https://doi.org/10.1207/s15327574ijt0403_1

Caseras, X., Garner, M., Bradley, B. P., & Mogg, K. (2007). Biases in visual orienting to negative and positive scenes in dysphoria: An eye movement study. *Journal of abnormal psychology*, 116(3), 491-497.

Cattell, R. B. (1966). The screen test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cerel, J., Maple, M., van de Venne, J., Moore, M., Flaherty, C., & Brown, M. (2016). Exposure to suicide in the community: Prevalence and correlates in one U.S. state. *Public Health Reports*, *131*(1), 100–107. https://doi.org/10.1177/003335491613100116

Chao, C.-M. (2019). Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.01652

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Cheah, J.-H., Thurasamy, R., Memon, M. A., Chuah, F., & Ting, H. (2020). Multigroup analysis using SmartPLS: Step-by-step guidelines for business research. *Asian Journal of Business Research*, *10*(3). https://doi.org/10.14707/ajbr.200087

Choi, E. P., Hui, B. P., & Wan, E. Y. (2020). Depression and anxiety in Hong Kong during COVID-19. *International Journal of Environmental Research and Public Health*, *17*(10), 3740. https://doi.org/10.3390/ijerph17103740

Clark, D. A., Beck, A. T., Alford, B. A., Bieling, P. J., & Segal, Z. V. (2000). Scientific Foundations of cognitive theory and therapy of depression. *Journal of Cognitive Psychotherapy*, *14*(1), 100–106. https://doi.org/10.1891/0889-8391.14.1.100

Clark, D. M., Wells, A. (1995). *A cognitive model of social phobia*. In R. Heimberg, M. Liebowitz, D. A. Hope, & F. R. Schneier (Eds.), Social phobia: Diagnosis, assessment, and treatment. New York: Guilford Press.

Coughlan, M., Cronin, P., & Ryan, F. (2007). Step-by-step guide to critiquing research. part 1: Quantitative research. *British Journal of Nursing*, 16(11), 658–663. https://doi.org/10.12968/bjon.2007.16.11.23681

Cowden Hindash, A. H., & Amir, N. (2011). Negative interpretation bias in individuals with depressive symptoms. *Cognitive Therapy and Research*, *36*(5), 502–511. https://doi.org/10.1007/s10608-011-9397-4

Cowden Hindash, A. H., & Rottenberg, J. (2015). Turning quickly on myself: Automatic interpretation biases in dysphoria are self-referent. *Cognition and Emotion*, *31*(2), 395–402. https://doi.org/10.1080/02699931.2015.1105792

Cowden Hindash, A. H., & Rottenberg, J. A. (2017). Moving towards the benign: Automatic interpretation bias modification in dysphoria. *Behavior Research and Therapy*, *99*, 98–107. https://doi.org/10.1016/j.brat.2017.09.005

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment*, Research & Evaluation, 10(7), 1-9.

Creswell, J.W., & Plano Clark, V.L. (2007). Designing and conducting mixed methods research. *Thousand Oaks*, CA: Sage Publications

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. https://doi.org/10.3758/brm.40.2.428

Dapprich, A. L., Lange, W.-G., Cima, M., & Becker, E. S. (2022). A validation of an ambiguous social scenario task for socially anxious and socially callous interpretations. *Cognitive Therapy and Research*, *46*(3), 608–619. https://doi.org/10.1007/s10608-021-10283-9

David, O. A., & David, D. (2022). How can we best use technology to teach children to regulate emotions? efficacy of the cognitive reappraisal strategy based on robot versus cartoons versus written statements in regulating test anxiety. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *40*(4), 793–802. https://doi.org/10.1007/s10942-021-00440-0

Davis, F.D. (1986) A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Sloan School of Management, Massachusetts Institute of Technology.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of Information Technology. *MIS Quarterly*, 13(3), 319. https://doi.org/10.2307/249008

Davranche, K., Tempest, G. D., Gajdos, T., & Radel, R. (2018). Impact of physical and cognitive exertion on cognitive control. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02369

Daştan, İ., & Gürler, C. (2016). Factors affecting the adoption of mobile payment systems: An empirical analysis. *EMAJ: Emerging Markets Journal*, 6(1), 17–24. https://doi.org/10.5195/emaj.2016.95

Dearing, K. F., & Gotlib, I. H. (2008). Interpretation of ambiguous information in girls at risk for depression. *Journal of Abnormal Child Psychology*, *37*(1), 79–91. https://doi.org/10.1007/s10802-008-9259-z

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology*, 1(20), 416-436.

Dimitrov, D. & Rumrill, P. (2003). Pretest-Posttest Designs and Measurement of Change. Work (Reading, Mass.). 20. 159-65.

Dishaw, M. T., & Strong, D. M. (1999). Extending the technology acceptance model with Task–Technology Fit Constructs. *Information & Management*, *36*(1), 9–21. https://doi.org/10.1016/s0378-7206(98)00101-3

Duque, S. I., Arnold, W. D., Odermatt, P., ， , X., Porensky, P. N., Schmelzer, L., Meyer, K., Kolb, S. J., Schümperli, D., Kaspar, B. K., & Burghes, A. H. (2015). A large animal model of spinal muscular atrophy and correction of phenotype. *Annals of Neurology*, *77*(3), 399–414. https://doi.org/10.1002/ana.24332

Edwards, C., Beattie, A. J., Edwards, A., & Spence, P. R. (2016). Differences in perceptions of communication quality between a Twitterbot and human agent for information seeking and learning. *Computers in Human Behavior*, *65*, 666–671. https://doi.org/10.1016/j.chb.2016.07.003

Elliott, R., Zahn, R., Deakin, J. F., & Anderson, I. M. (2010). Affective cognition and its disruption in mood disorders. *Neuropsychopharmacology*, *36*(1), 153–182. https://doi.org/10.1038/npp.2010.77

Erford, B. T., Johnson, E., & Bardoshi, G. (2016). Meta-analysis of the English version of the Beck Depression Inventory–Second Edition. *Measurement and Evaluation in Counseling and Development*, *49*(1), 3–33. https://doi.org/10.1177/0748175615596783

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1. https://doi.org/10.11648/j.ajtas.20160501.11

Erevelles, S., & Leavitt, C. (1992). A comparison of current models of consumer satisfaction/ dissatisfaction. *Journal of Consumer Satisfaction, Dissatisfaction, and Complaining Behavior*, 5(104).

Everaert, J., & Koster, E. H. W. (2020). The interplay among attention, interpretation, and memory biases in depression: Revisiting the combined cognitive bias hypothesis. *Cognitive Biases in Health and Psychiatric Disorders*, 193–213. https://doi.org/10.1016/b978-0-12-816660-4.00009-x

Everaert, J., Duyck, W., & Koster, E. H. (2014). Attention, interpretation, and memory biases in subclinical depression: A proof-of-principle test of the combined cognitive biases hypothesis. *Emotion*, *14*(2), 331–340. https://doi.org/10.1037/a0035250

Everaert, J., Koster, E. H. W., & Derakshan, N. (2012). The combined cognitive bias hypothesis in depression. *Clinical Psychology Review*, *32*(5), 413–424. https://doi.org/10.1016/j.cpr.2012.04.003

Everaert, J., Tierens, M., Uzieblo, K., & Koster, E. H. (2013). The indirect effect of attention bias on memory via interpretation bias: Evidence for the combined cognitive bias hypothesis in subclinical depression. *Cognition & Emotion*, *27*(8), 1450–1459. https://doi.org/10.1080/02699931.2013.787972

Eysenck, M. W., Mogg, K., May, J., Richards, A., & Mathews, A. (1991). Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of abnormal psychology*, 100(2), 144-150.

Feng, Y.-C., Krahé, C., Meeten, F., Sumich, A., Mok, C. L. M., & Hirsch, C. R. (2020). Impact of imagery-enhanced interpretation training on offline and online interpretations in worry. *Behavior Research and Therapy*, *124*, 103497. https://doi.org/10.1016/j.brat.2019.103497

Feng, J., & Sears, A. (2009). Speech input to support universal access. In *The Universal Access Handbook* (Vol. 30, pp. 1–16). essay, CRC Press.

Field, A. (2014). *Discovering statistics using Ibm Spss statistics.* SAGE Publications.

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., &Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behavior Research and Therapy*, 116, 69–79. https://doi.org/10.1016/j.brat.2019.01.010

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2). https://doi.org/10.2196/mental.7785

Forbes, E. E., & Dahl, R. E. (2011). Research review: Altered reward function in adolescent depression: What, when and how? *Journal of Child Psychology and Psychiatry*, *53*(1), 3–15. https://doi.org/10.1111/j.1469-7610.2011.02477.x

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, Artificial Intelligence (AI), and the nature of autonomy. *Minds and Machines*, *31*(4), 595–616. https://doi.org/10.1007/s11023-021-09579-2

Fosch-Villaronga, E., & Drukarch, H. (2022). AI for healthcare robotics. CRC Press

Fox, W. S., & Denzin, N. K. (1979). The research act: A theoretical introduction to sociological methods. *Contemporary Sociology*, *8*(5), 750. https://doi.org/10.2307/2065439

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (TESS) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health*, *5*(4). https://doi.org/10.2196/mental.9782

García de Blanes Sebastián, M., Sarmiento Guede, J. R., & Antonovica, A. (2022). Application and extension of the UTAUT2 model for determining behavioral intention factors in use of the Artificial Intelligence Virtual assistants. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.993935

Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2019). Assessing the effect of persuasive robots' interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. *Advanced Robotics*, *33*(7-8), 325–337. https://doi.org/10.1080/01691864.2019.1589570

Giedd, J. N. (2004). Structural magnetic resonance imaging of the adolescent brain. *Annals of the New York Academy of Sciences*, 1021(1), 77–85. https://doi.org/10.1196/annals.1308.009

Girden, E. R. (2003). *ANOVA: Repeated measures*. Sage Publications.

Glaser, B. G., & Strauss, A. L. (2017). *The discovery of Grounded Theory: Strategies for qualitative research*. Routledge.

Glenn, D., Golinelli, D., Rose, R. D., Roy-Byrne, P., Stein, M. B., Sullivan, G., Bystritksy, A., Sherbourne, C., & Craske, M. G. (2013). Who gets the most out of cognitive behavioral

therapy for anxiety disorders? the role of treatment dose and patient engagement. *Journal of Consulting and Clinical Psychology*, 81(4), 639–649. https://doi.org/10.1037/a0033403

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Gonsalves, M., Whittles, R. L., Weisberg, R. B., & Beard, C. (2019). A systematic review of the word sentence association paradigm (WSAP). *Journal of Behavior Therapy and Experimental Psychiatry*, *64*, 133–148. https://doi.org/10.1016/j.jbtep.2019.04.003

Goodhue, D. L. (1998). Development and measurement validity of a task-technology fit instrument for user evaluations of information system. *Decision Sciences*, *29*(1), 105–138. https://doi.org/10.1111/j.1540-5915.1998.tb01346.x

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, *19*(2), 213. https://doi.org/10.2307/249689

Görgen, S. M. (2015). The role of mental imagery in depression: Negative mental imagery induces strong implicit and explicit affect in depression. *Frontiers in Psychiatry*, *6*. https://doi.org/10.3389/fpsyt.2015.00094

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and Future Directions. *Annual Review of Clinical Psychology*, *6*(1), 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Gotlib, I. H., & Krasnoperova, E. (1998). Biased Information Processing as a vulnerability factor for depression. *Behavior Therapy*, *29*(4), 603–617. https://doi.org/10.1016/s0005-7894(98)80020-8

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, 49, 157–169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Gaver, W. W. (1991). Technology affordances. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology* - CHI '91. https://doi.org/10.1145/108844.108856

Greene, J. C. (2008). Is Mixed Methods Social Inquiry a Distinctive Methodology? Journal of Mixed Methods Research, 2(1), 7-22. https://doi.org/10.1177/1558689807309969

Guemghar, I., Pires de Oliveira Padilha, P., Abdel-Baki, A., Jutras-Aswad, D., Paquette, J., & Pomey, M.-P. (2022). Social Robot Interventions in mental health care and their outcomes, barriers, and facilitators: Scoping review. *JMIR Mental Health*, *9*(4). https://doi.org/10.2196/36094

Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2018). *Multivariate Data Analysis: Eighth Edition*, United Kingdom, Cengage Learning EMEA.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., Schaefer, K. E., & Szalma, J. L. (2023). How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1081086

Harrison, Steve & Tatar, Deborah & Sengers, Phoebe. (2007). The three paradigms of HCI.

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). *Mental Health Chatbot for Young Adults with Depression Symptoms during the COVID-19 Pandemic: A Single-Blind, Three-Arm, Randomized Controlled Trial (Preprint)*. https://doi.org/10.2196/preprints.40719

Henricks, L. A., Lange, W.-G., Luijten, M., & Becker, E. S. (2022). A new social picture task to assess interpretation bias related to social fears in adolescents. *Research on Child and Adolescent Psychopathology*. https://doi.org/10.1007/s10802-022-00915-3

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, 33(3), 405–431. https://doi.org/10.1108/imr-09-2014-0304

Hess, T. M., Germain, C. M., Swaim, E. L., & Osowski, N. L. (2009). Aging and selective engagement: The moderating impact of motivation on older adults' resource utilization. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *64B* (4), 447–456. https://doi.org/10.1093/geronb/gbp020

Hess, T. M., Queen, T. L., & Ennis, G. E. (2012). Age and self-relevance effects on information search during decision making. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(5), 703–711. https://doi.org/10.1093/geronb/gbs108

Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, *21*(1), 100–120. https://doi.org/10.1177/109634809702100108

Hirsch, C. R., Krahé, C., Whyte, J., Bridge, L., Loizou, S., Norton, S., & Mathews, A. (2020). Effects of modifying interpretation bias on transdiagnostic repetitive negative thinking. *Journal of Consulting and Clinical Psychology*, 88(3), 226–239. https://doi.org/10.1037/ccp0000455

Hirsch, C. R., Krahé, C., Whyte, J., Loizou, S., Bridge, L., Norton, S., & Mathews, A. (2018). Interpretation training to target repetitive negative thinking in generalized anxiety disorder and depression. *Journal of Consulting and Clinical Psychology*, 86(12), 1017–1030. https://doi.org/10.1037/ccp0000310

Ho, H. M. (2017). *The lived experience of foreign domestic helpers in caring for older people in the community: a hermeneutic phenomenological study*. Hong Kong: School of Nursing, The Hong Kong Polytechnic University

Hofmann, S. G. (2004). Cognitive mediation of treatment change in social phobia. *Journal of Consulting and Clinical Psychology*, 72(3), 392–399. https://doi.org/10.1037/0022-006x.72.3.392

Hohensee, N., Meyer, M. J., & Teachman, B. A. (2020). The effect of confidence on dropout rate and outcomes in online cognitive bias modification. *Journal of Technology in Behavioral Science*, *5*(3), 226–234. https://doi.org/10.1007/s41347-020-00129-8

Hollon, S. D. (2019). Cognitive behavior therapy. *Depression*, 271–286. https://doi.org/10.1093/med/9780190929565.003.0016

Holmes, D. S. (1973). Effectiveness of debriefing after a stress-producing deception. *Journal of Research in Personality*, 7(2), 127–138. https://doi.org/10.1016/0092-6566(73)90046-9

Holmes, E. A., & Mathews, A. (2005). Mental imagery and emotion: A special relationship? *Emotion*, *5*(4), 489–497. https://doi.org/10.1037/1528-3542.5.4.489

Holmes, E. A., Lang, T. J., & Shah, D. M. (2009). Developing interpretation bias modification as a "cognitive vaccine" for depressed mood: Imagining positive events makes you feel better than thinking about them verbally. *Journal of Abnormal Psychology*, *118*(1), 76–88. https://doi.org/10.1037/a0012590

Holmes, E. A., Mathews, A., Dalgleish, T., & Mackintosh, B. (2006). Positive interpretation training: Effects of mental imagery versus verbal training on positive mood. *Behavior Therapy*, *37*(3), 237–247. https://doi.org/10.1016/j.beth.2006.02.002

Holmes, E. A., Mathews, A., Mackintosh, B., & Dalgleish, T. (2008). The causal effect of mental imagery on emotion assessed using picture-word cues. *Emotion*, *8*(3), 395–409. https://doi.org/10.1037/1528-3542.8.3.395

Hoorn, J. F. (2015a). Machine Medical Ethics: When a human is delusive but the machine has its wits about him. *Machine Medical Ethics*, 233–254. https://doi.org/10.1007/978-3-319-08108-3_15

Hoorn, J. F. (2015b). Psychological aspects of technology interacting with humans. *The Handbook of the Psychology of Communication Technology*, 176–201. https://doi.org/10.1002/9781118426456.ch8

Hoorn, J. F. (2020a). Theory of robot communication: I. the medium is the Communication partner. *International Journal of Humanoid Robotics*, 17(06), 2050026. https://doi.org/10.1142/s0219843620500267

Hoorn, J. F. (2020b). Theory of robot communication: II. Befriending a robot over time. *International Journal of Humanoid Robotics*, 17(6), 2050027. doi: 10.1142/S0219843620500279

Hoorn, J. F., Baier, T., Van Maanen, J. A. N., & Wester, J. (2021/2023). Silicon Coppélia and the formalization of the affective process. *IEEE Transactions on Affective Computing*, 14(1), 255-278. doi: 10.1109/TAFFC.2020.3048587

Hoorn, J. F., & Huang, I. S. (2024). The media inequality, Uncanny Mountain, and the singularity is far from near: Iwaa and sophia robot versus a real human being. *International Journal of Human-Computer Studies*, *181*, 103142. https://doi.org/10.1016/j.ijhcs.2023.103142

Hoorn, J. F., & Winter, S. D. (2017). Here comes the bad news: Doctor robot taking over. *International Journal of Social Robotics*, *10*(4), 519–535. https://doi.org/10.1007/s12369-017-0455-2

Hoorn, J. F., Konijn, E. A., & Pontier, M. A. (2018). Dating a synthetic character is like dating a man. *International Journal of Social Robotics*, 1-19. doi: 10.1007/s12369-018-0496-1

Hoorn. J. F., & Huang, I. S. (2024). The Media Inequality, Uncanny Mountain, and the Singularity is far from near: Iwaa and Sophia robot versus a real human being. *International Journal of Human-Computer Studies*, 128, 103142. doi: 10.1016/j.ijhcs.2023.103142

Hoppitt, L., Illingworth, J. L., MacLeod, C., Hampshire, A., Dunn, B. D., & Mackintosh, B. (2014). Modifying social anxiety related to a real-life stressor using online cognitive bias modification for interpretation. *Behavior Research and Therapy*, *52*, 45–52. https://doi.org/10.1016/j.brat.2013.10.008

Hoppitt, L., Mathews, A., Yiend, J., & Mackintosh, B. (2010). Cognitive bias modification: The critical role of active training in modifying emotional responses. *Behavior Therapy*, *41*(1), 73–81. https://doi.org/10.1016/j.beth.2009.01.002

Hou, W. K., & Hall, B. J. (2019). The Mental Health Impact of the pro-democracy movement in Hong Kong. *The Lancet Psychiatry*, *6*(12), 982. https://doi.org/10.1016/s2215-0366(19)30382-7

Huang, I. S., Cheung, Y. W. Y., & Hoorn, J. F. (2023). Loving-kindness and walking meditation with a robot: Countering negative mood by stimulating creativity. *International Journal of Human-Computer Studies*, *179*, 103107. https://doi.org/10.1016/j.ijhcs.2023.103107

Huber, B., & Gajos, K. Z. (2020). Conducting online virtual environment experiments with uncompensated, unsupervised samples. *PLOS ONE*, *15*(1). https://doi.org/10.1371/journal.pone.0227629

Hult, G. T., Ketchen, D. J., Griffith, D. A., Finnegan, C. A., Gonzalez-Padron, T., Harmancioglu, N., Huang, Y., Talay, M. B., & Cavusgil, S. T. (2008). Data equivalence in cross-cultural International Business Research: Assessment and Guidelines. *Journal of International Business Studies*, 39(6), 1027–1044. https://doi.org/10.1057/palgrave.jibs.8400396

Ingram, R. E. (1984). Toward an information-processing analysis of depression. *Cognitive Therapy and Research*, *8*(5), 443–477. https://doi.org/10.1007/bf01173284

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (WYSA) for digital mental well-being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11). https://doi.org/10.2196/12106

Jackson, P. W., & Messick, S. (1965). The person, the product, and the response: Conceptual problems in the assessment of creativity1. *Journal of Personality*, *33*(3), 309–329. https://doi.org/10.1111/j.1467-6494.1965.tb01389.x

Javeline, D. (1999). Response effects in polite cultures: A test of acquiescence in Kazakhstan. *Public Opinion Quarterly*, *63*(1), 1. https://doi.org/10.1086/297701

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

Jennings, M. A., & Cribbie, R. A. (2021). Comparing pre-post change across groups: Guidelines for choosing between difference scores, ANCOVA, and residual change scores. *Journal of Data Science*, 14(2), 205–230. https://doi.org/10.6339/jds.201604_14(2).0002

Jin, M., Ji, L., & Peng, H. (2019). The relationship between cognitive abilities and the decision-making process: The moderating role of self-relevance. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01892

Johnson, R. & Onwuegbuzie, Anthony. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational researcher. 33. 14. 10.3102/0013189X033007014.

Joormann, J., Waugh, C. E., & Gotlib, I. H. (2015). Cognitive bias modification for interpretation in major depression. *Clinical Psychological Science*, *3*(1), 126–139. https://doi.org/10.1177/2167702614560748

Jøranson, N., Pedersen, I., Rokstad, A. M., & Ihlebæk, C. (2016). Change in quality of life in older people with dementia participating in paro-activity: A cluster-randomized controlled trial. *Journal of Advanced Nursing*, *72*(12), 3020–3033. https://doi.org/10.1111/jan.13076

Jung, M. M., van der Leij, L., & Kelders, S. M. (2017). An exploration of the benefits of an animallike robot companion with more advanced touch interaction capabilities for Dementia Care. *Frontiers in ICT*, *4*. https://doi.org/10.3389/fict.2017.00016

Kaiser, H. F. (1960). The application of electronic computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1974). An index of Factorial Simplicity. *Psychometrika*, 39(1), 31–36. https://doi.org/10.1007/bf02291575

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., Mackinnon, A., Meyer, B., Botella, C., Littlewood, E., Andersson, G., Christensen, H., Klein, J. P., Schröder, J., Bretón-López, J., Scheider, J., Griffiths, K., Farrer, L., Huibers, M. J., … Cuijpers, P. (2017). Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of

depressive symptoms. *JAMA Psychiatry*, *74*(4), 351. https://doi.org/10.1001/jamapsychiatry.2017.0044

Karim, R., Lopez, E., Oleson, K., Li, T., Björling, E. A., & Cakmak, M. (2022). Share with me: A study on a social robot collecting mental health data. *Lecture Notes in Computer Science*, 218–227. https://doi.org/10.1007/978-3-031-24667-8_20

Kawamichi, H., Sugawara, S. K., Hamano, Y. H., Makita, K., Kochiyama, T., & Sadato, N. (2016). Increased frequency of social interaction is associated with enjoyment enhancement and reward system activation. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep24561

Keller, A. S., Leikauf, J. E., Holt-Gosselin, B., Staveland, B. R., & Williams, L. M. (2019). Paying attention to attention in depression. *Translational psychiatry*, 9(1), 1-12.

Kellough, J. L., Beevers, C. G., Ellis, A. J., & Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: An eye tracking study. *Behavior Research and Therapy*, *46*(11), 1238–1243. https://doi.org/10.1016/j.brat.2008.07.004

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, 77, 101925. https://doi.org/10.1016/j.tele.2022.101925

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & ??st??n, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry*, *20*(4), 359–364. https://doi.org/10.1097/yco.0b013e32816ebc8c

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 593. https://doi.org/10.1001/archpsyc.62.6.593

Kevin, B., VIkin, B., & Nair, M. (2023). *Building a Chatbot for Healthcare Using NLP*. https://doi.org/10.36227/techrxiv.22578472.v1

Khalaf, A. M., Alubied, A. A., Khalaf, A. M., & Rifaey, A. A. (2023). The impact of social media on the mental health of adolescents and young adults: A systematic review. *Cureus*. https://doi.org/10.7759/cureus.42990

Khawaja, Z., & Bélisle-Pipon, J.-C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-Powered Mental Health Chatbots. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1278186

Kirk, R. (2013). *Experimental Design: Procedures for the Behavioral Sciences*. https://doi.org/10.4135/9781483384733

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2012). A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*, 73(1), 47–62. https://doi.org/10.1177/0013164412450574

Kline, P. (2016). *A Handbook of Test Construction: Introduction to psychometric design*. Routledge.

Klüber, K., & Onnasch, L. (2022). When Robots Fail—a VR investigation on caregivers' tolerance towards communication and processing failures. *Robotics*, 11(5), 106. https://doi.org/10.3390/robotics11050106

Koster, E. H. W., Hoorelbeke, K., Onraedt, T., Owens, M., & Derakshan, N. (2017). Cognitive control interventions for depression: A systematic review of findings from Training Studies. *Clinical Psychology Review*, *53*, 79–92. https://doi.org/10.1016/j.cpr.2017.02.002

Krawietz, S. A., Tamplin, A. K., & Radvansky, G. A. (2012). Aging and mind wandering during text comprehension. *Psychology and Aging*, *27*(4), 951–958. https://doi.org/10.1037/a0028831

Krahé, C., Whyte, J., Bridge, L., Loizou, S., & Hirsch, C. R. (2019). Are different forms of repetitive negative thinking associated with interpretation bias in generalized anxiety disorder and depression? *Clinical Psychological Science*, 7(5), 969–981. https://doi.org/10.1177/2167702619851808

Kuiper, N. A., Olinger, L. J., & Martin, R. A. (1988). Dysfunctional attitudes, stress, and negative emotions. *Cognitive Therapy and Research*, *12*(6), 533–547. https://doi.org/10.1007/bf01205008

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, *69*, 313–332. https://doi.org/10.1016/j.neubiorev.2016.07.002

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/s1364-6613(00)01560-6

Lai, A., & Tetreault, J. (2018). Discourse coherence in the wild: A dataset, evaluation and methods. *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. https://doi.org/10.18653/v1/w18-5023

Lam, M. I., Cai, H., Chen, P., Lok, K.-I., Chow, I. H., Si, T. L., Su, Z., Ng, C. H., An, F.-R., & Xiang, Y.-T. (2024). The inter-relationships between depressive symptoms and suicidality among Macau residents after the "Relatively static management" COVID-19 strategy: A perspective of network analysis. *Neuropsychiatric Disease and Treatment, Volume 20*, 195–209. https://doi.org/10.2147/ndt.s451031

Lam, R. W., Kennedy, S. H., McIntyre, R. S., & Khullar, A. (2014). Cognitive dysfunction in major depressive disorder: Effects on psychosocial functioning and implications for treatment. *The Canadian Journal of Psychiatry*, *59*(12), 649–654. https://doi.org/10.1177/070674371405901206

Lang, T. J., Blackwell, S. E., Harmer, C. J., Davison, P., & Holmes, E. A. (2012). Cognitive bias modification using mental imagery for depression: Developing a novel computerized intervention to change negative thinking styles. *European Journal of Personality*, *26*(2), 145–157. https://doi.org/10.1002/per.855

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*. https://doi.org/10.3115/1075096.1075165

Law, M., Jarrett, P., Nieuwoudt, M. K., Holtkamp, H., Giglio, C., & Broadbent, E. (2022). The effects of interacting with a paro robot after a stressor in patients with psoriasis: A randomised pilot study. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.871295

Lazar, J., Feng, J.H. & Hochheiser, H. (2010). *Research methods in human-computer interaction*. Chichester: John Wiley.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer.

Lee, I., & Yang, J. (2009). Common clustering algorithms. *Comprehensive Chemometrics*, 577–618. https://doi.org/10.1016/b978-044452701-1.00064-8

Lee, D., & Daunizeau, J. (2019). Choosing what we like vs liking what we choose: How choice-induced preference change might actually be instrumental to decision-making. https://doi.org/10.1101/661116

Letchumanan, M., & Tarmizi, R. (2011). Assessing the intention to use e-book among engineering undergraduates in Universiti Putra Malaysia, Malaysia. *Library Hi Tech*, *29*(3), 512–528. https://doi.org/10.1108/07378831111174459

Levitt, S & List, J. (2005). What Do Laboratory Experiments Tell Us About the Real World?. *Journal of Economic Perspectives*. 21.

Li, H., Zhang, R., Lee, Y.-C., Kraut, R., & Mohr, D. C. (2023). *Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being.* https://doi.org/10.31234/osf.io/m3vjt

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Li, Y., Welbourne, E., & Landay, J. A. (2006). Design and experimental analysis of continuous location tracking techniques for Wizard of Oz Testing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/1124772.1124924

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77, 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Liu, X.-Q., Guo, Y.-X., Zhang, W.-J., & Gao, W.-J. (2022). Influencing factors, prediction and prevention of depression in college students: A literature review. *World Journal of Psychiatry*, *12*(7), 860–873. https://doi.org/10.5498/wjp.v12.i7.860

Lopes, S. L., Ferreira, A. I., & Prada, R. (2023). The use of robots in the workplace: Conclusions from a health promoting intervention using social robots. *International Journal of Social Robotics*, 15(6), 893–905. https://doi.org/10.1007/s12369-023-01000-5

Lopez, A., Ccasane, B., Paredes, R., & Cuellar, F. (2017). Effects of using indirect language by a robot to change human attitudes. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* https://doi.org/10.1145/3029798.3038310

Lu, M.-L., Che, H., Chang, S., & Shen, W. W. (2002). Reliability and validity of the Chinese version of the Beck Depression Inventory-II. *Chinese Mental Health Journal*, 25(6), 476–480.

Lu, Mong-Liang & Che, HH & Chang, Shangwen & Shen, W.W.. (2002). Reliability and Validity of the Chinese Version of the Beck Depression Inventory-II. *Taiwanese Journal of Psychiatry*. 16. 301-310.

Luo, R., Zhang, T., Chen, D., Hoorn, J. F., & Huang, I. (2022). *Social Robots Outdo the Not-so-Social Media for Self-Disclosure: Safe Machines Preferred to Unsafe Humans?* https://doi.org/10.20944/preprints202206.0233.v1

Ly, K. H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*, 10, 39–46. https://doi.org/10.1016/j.invent.2017.10.002

Mackay, W. E., & Fayard, A.-L. (1997). HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines. *Proceedings of the Conference on Designing Interactive Systems Processes, Practices, Methods, and Techniques - DIS '97*. https://doi.org/10.1145/263552.263612

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (2006). An integrative model of Organizational Trust. Organizational Trust, 82–108. https://doi.org/10.1093/oso/9780199288496.003.0004

Martin, B., & Hanington, B. M. (2012). *Universal Methods of Design: 100 Ways To Research Complex Problems, develop innovative ideas, and Design Effective Solutions*. Rockport Publishers.

Mathews, A. (2012). Effects of modifying the interpretation of emotional ambiguity. *Journal of Cognitive Psychology*, *24*(1), 92–105. https://doi.org/10.1080/20445911.2011.584527

Mathews, A., & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of Abnormal Psychology*, *109*(4), 602–615. https://doi.org/10.1037/0021-843x.109.4.602

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology*, *1*(1), 167–195. https://doi.org/10.1146/annurev.clinpsy.1.102803.143916

Mathews, A., Richards, A., & Eysenck, M. (1989). Interpretation of homophones related to threat in anxiety states. *Journal of Abnormal Psychology*, *98*(1), 31–34. https://doi.org/10.1037/0021-843x.98.1.31

Mathews, A., Ridgeway, V., & Williamson, D. A. (1996). Evidence for attention to threatening stimuli in depression. *Behavior research and therapy*, 34(9), 695-705.

Mattick, R. P., Peters, L., & Clarke, J. C. (1989). Exposure and cognitive restructuring for social phobia: A controlled study. *Behavior Therapy*, 20(1), 3–23. https://doi.org/10.1016/s0005-7894(89)80115-7

McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics*, 31(4), 1–11. https://doi.org/10.1145/2185520.2335442

McHugh, M. L. (2012). Interrater Reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

McGillivray, J. A., & Kershaw, M. (2013). Do we need both cognitive and behavioral components in interventions for depressed mood in people with mild intellectual disability? *Journal of Intellectual Disability Research*, 59(2), 105–115. https://doi.org/10.1111/jir.12110

Mertler C. and Reinhart, R. V. (2016). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation: Sixtth Edition*, United States, Taylor and Francis.

Micco, J. A., Henin, A., & Hirshfeld-Becker, D. R. (2013). Efficacy of interpretation bias modification in depressed adolescents and young adults. *Cognitive Therapy and Research*, 38(2), 89–102. https://doi.org/10.1007/s10608-013-9578-4

Mira, A., Bretón-López, J., García-Palacios, A., Quero, S., Baños, R. M., & Botella, C. (2017). An internet-based program for depressive symptoms using human and automated support: A

randomized controlled trial. *Neuropsychiatric Disease and Treatment*, *Volume 13*, 987–1006. https://doi.org/10.2147/ndt.s130994

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4. https://doi.org/10.3389/frobt.2017.00021

Miyamoto, S., Abe, R., Endo, Y., & Takeshita, J. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*. https://doi.org/10.1109/socpar.2015.7492784

Moerman, C. J., van der Heide, L., & Heerink, M. (2018). Social Robots to support children's well-being under medical treatment: A systematic state-of-the-art review. *Journal of Child Health Care*, *23*(4), 596–612. https://doi.org/10.1177/1367493518803031

Mollahosseini, A., Abdollahi, H., Sweeny, T. D., Cole, R., & Mahoor, M. H. (2018). Role of embodiment and presence in human perception of robots' facial cues. *International Journal of Human-Computer Studies*, *116*, 25–39. https://doi.org/10.1016/j.ijhcs.2018.04.005

Molli, V. L. P. (2022). Effectiveness of AI-Based Chatbots in Mental Health Support: A Systematic Review. Journal of Healthcare AI and ML , 9(9), 1-11.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological bulletin*, 132(2), 297-326

Morgan, D. L. (2007). Paradigms lost and Pragmatism regained. *Journal of Mixed Methods Research*, *1*(1), 48–76. https://doi.org/10.1177/2345678906292462

Morii, M., Sakagami, T., Masuda, S., Okubo, S., & Tamari, Y. (2017). Correction to: How does response bias emerge in lengthy sequential preference judgments? *Behaviormetrika*. https://doi.org/10.1007/s41237-017-0044-6

Morina, N., Deeprose, C., Pusowski, C., Schmid, M., & Holmes, E. A. (2011). Prospective mental imagery in patients with major depressive disorder or anxiety disorders. *Journal of Anxiety Disorders*, *25*(8), 1032–1037. https://doi.org/10.1016/j.janxdis.2011.06.012

Morse M. J. & Niehaus L. (2009). mixed method design: Principles and procedures. Walnut Creek, CA, USA: Left Coast Press Inc.; 193 pages; ISBN 978-1-59874-298-5

Mouloudj, K., Bouarar, A. C., Asanza, D. M., Saadaoui, L., Mouloudj, S., Njoku, A. U., Evans, M. A., & Bouarar, A. (2023). Factors influencing the adoption of digital health apps. *Advances in Healthcare Information Systems and Administration*, 116–132. https://doi.org/10.4018/978-1-6684-8337-4.ch007

Moyle, W., Bramble, M., Jones, C., & Murfield, J. (2016). Care staff perceptions of a social robot called Paro and a look-alike plush toy: A descriptive qualitative approach. *Aging & Mental Health*, *22*(3), 330–335. https://doi.org/10.1080/13607863.2016.1262820

Mueller, E. M., Pechtel, P., Cohen, A. L., Douglas, S. R., & Pizzagalli, D. A. (2015). Potentiated processing of negative feedback in depression is attenuated by anhedonia. *Depression and Anxiety*, 32(4), 296–305. https://doi.org/10.1002/da.22338

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction.* https://doi.org/10.1145/1514095.1514110

Ng, A. P., Chin, W. Y., Wan, E. Y., Chen, J., & Lau, C. S. (2021). Prevalence of depression and suicide ideation in Hong Kong doctors: A cross-sectional study. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-98668-4

Ni, M. Y., Yao, X. I., Leung, K. S., Yau, C., Leung, C. M., Lun, P., Flores, F. P., Chang, W. C., Cowling, B. J., & Leung, G. M. (2020). Depression and post-traumatic stress during major social unrest in Hong Kong: A 10-year prospective Cohort Study. *The Lancet*, *395*(10220), 273–284. https://doi.org/10.1016/s0140-6736(19)33160-5

Nicol, G., Wang, R., Graham, S., Dodd, S., & Garbutt, J. (2022). Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: Feasibility and acceptability study. *JMIR Formative Research*, 6(11). https://doi.org/10.2196/40242

O'Connor, C. E., Everaert, J., & Fitzgerald, A. (2021). Interpreting ambiguous emotional information: Convergence among interpretation bias measures and unique relations with depression severity. *Journal of Clinical Psychology*, 77(11), 2529–2544. https://doi.org/10.1002/jclp.23186

O'Grady, M. A., Tennen, H., & Armeli, S. (2010). Depression history, depression vulnerability, and the experience of everyday negative events. *Journal of Social and Clinical Psychology*, 29(9), 949–974. https://doi.org/10.1521/jscp.2010.29.9.949

O'Neil, K. M., Penrod, S. D., & Bornstein, B. H. (2003). Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 217–226. https://doi.org/10.3758/bf03202544

Ogawa, M., Oyama, G., Morito, K., Kobayashi, M., Yamada, Y., Shinkawa, K., Kamo, H., Hatano, T., & Hattori, N. (2022). Can ai make people happy? the effect of AI-based chatbot on smile and speech in parkinson's disease. *Parkinsonism & Related Disorders*, *99*, 43–46. https://doi.org/10.1016/j.parkreldis.2022.04.018

Oliver, R. L., & DeSarbo, W. S. (1988). Response determinants in satisfaction judgments. *Journal of Consumer Research*, *14*(4), 495. https://doi.org/10.1086/209131

Oliver, R. L. (1997). *Satisfaction: A behavioral perspective on the consumer*. McGraw-Hill.

Ouimet, A. J., Gawronski, B., & Dozois, D. J. A. (2009). Cognitive vulnerability to anxiety: A review and an integrative model. *Clinical Psychology Review*, *29*(6), 459–470. https://doi.org/10.1016/j.cpr.2009.05.004

Paulhus, D. L. (1991). Measurement and control of response bias. *Measures of Personality and Social Psychological Attitudes*, 17–59. https://doi.org/10.1016/b978-0-12-590241-0.50006-x

Paulo, M. M., Rita, P., Oliveira, T., & Moro, S. (2018). Understanding mobile augmented reality adoption in a consumer context. *Journal of Hospitality and Tourism Technology*, 9(2), 142–157. https://doi.org/10.1108/jhtt-01-2017-0006

Pearson, J., & Kosslyn, S. M. (2013). Mental imagery. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00198

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences*, *19*(10), 590–602. https://doi.org/10.1016/j.tics.2015.08.003

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. *arXiv preprint arXiv:1909.01066*.

Peyton, K., Huber, G. A., & Coppock, A. (2020). *The Generalizability of Online Experiments Conducted during the COVID-19 Pandemic*. https://doi.org/10.31235/osf.io/s45yg

Phaosathianphan, N., & Leelasantitham, A. (2019). Understanding the adoption factors influence on the use of Intelligent Travel assistant (ITA) for eco-tourists: An extension of the utaut. *International Journal of Innovation and Technology Management*, 16(08). https://doi.org/10.1142/s0219877019500603

Pollak, C., Wexler, S. S., & Drury, L. (2022). Effect of a robotic pet on social and physical frailty in community-dwelling older adults: A randomized controlled trial. *Research in Gerontological Nursing*, *15*(5), 229–237. https://doi.org/10.3928/19404921-20220830-01

Ponte, K. (2022, January 10). *Understanding Mental Illness Triggers*. NAMI. https://www.nami.org/recovery/understanding-mental-illness-triggers/

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/1228716.1228736

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., & Choi, Y. (2018). Event2Mind: Commonsense inference on events, intents, and reactions. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p18-1043

Ratcliff, R., Huang-Pollock, C., & McKoon, G. (2018). Modeling individual differences in the go/no-go task with a diffusion model. *Decision*, *5*(1), 42–62. https://doi.org/10.1037/dec0000065

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.1026//1618-3169.49.4.243

Robinson, H., MacDonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: A randomized controlled trial. *Journal of the American Medical Directors Association*, *14*(9), 661–667. https://doi.org/10.1016/j.jamda.2013.02.007

Robinson, N. L., Cottier, T. V., & Kavanagh, D. J. (2019). Psychosocial Health Interventions by Social Robots: Systematic review of Randomized Controlled Trials. *Journal of Medical Internet Research*, *21*(5). https://doi.org/10.2196/13203

Rohrbacher, H., & Reinecke, A. (2014). Measuring change in depression-related interpretation bias: Development and validation of a parallel ambiguous scenarios test. *Cognitive Behavior Therapy*, *43*(3), 239–250. https://doi.org/10.1080/16506073.2014.919605

Rohrbacher, H., Blackwell, S. E., Holmes, E. A., & Reinecke, A. (2014). Optimizing the ingredients for imagery-based interpretation bias modification for depressed mood: Is self-generation more effective than imagination alone? *Journal of Affective Disorders*, *152-154*, 212–218. https://doi.org/10.1016/j.jad.2013.09.013

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2023). A matter of consequences Understanding the effects of robot errors on people's trust in HRI. *Interaction Studies*, 24(3), 380–421. https://doi.org/10.1075/is.21025.ros

Rude, S. S., Valdez, C. R., Odom, S., & Ebrahimi, A. (2003). Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research*, 27(4), 415–429. https://doi.org/10.1023/a:1025472413805

Russo, S. J., & Nestler, E. J. (2013). The brain reward circuitry in mood disorders. *Nature Reviews Neuroscience*, *14*(9), 609–625. https://doi.org/10.1038/nrn3381

Sabanovic, S., Bennett, C. C., Wan-Ling Chang, & Huber, L. (2013). Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. https://doi.org/10.1109/icorr.2013.6650427

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment*, *11*(2–3), 194–205. https://doi.org/10.1111/1468-2389.00243

Salgado, T., Tavares, J., & Oliveira, T. (2020). Drivers of Mobile Health acceptance and use from the patient perspective: Survey study and quantitative model development. *JMIR mHealth and uHealth*, *8*(7). https://doi.org/10.2196/17588

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2696454.2696497

Samuels, P. (2017). Advice on Reliability Analysis with Small Samples - Revised Version.

Sandoval, E. B., Mubin, O., & Obaid, M. (2014). Human robot interaction and fiction: A contradiction. *Social Robotics*, 54–63. https://doi.org/10.1007/978-3-319-11973-1_6

Sanchez, A., Vazquez, C., Marker, C., LeMoult, J., & Joormann, J. (2013). Attentional disengagement predicts stress recovery in depression: An eye-tracking study. *Journal of Abnormal Psychology*, 122(2), 303-313.

Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and Natural Language Inference. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. https://doi.org/10.18653/v1/2021.eacl-main.20

Schick, A., Feine, J., Morana, S., Maedche, A., & Reininghaus, U. (2022). Validity of chatbot use for Mental Health Assessment: Experimental Study. *JMIR mHealth and uHealth*, *10*(10). https://doi.org/10.2196/28082

Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *69*(S2), 107–131. https://doi.org/10.1007/s11577-017-0454-1

Schramm, L. T., Dufault, D., & Young, J. E. (2020). Warning: This robot is not what it seems! exploring expectation discrepancy resulting from robot design. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3371382.3378280

Schweizer, S., Hampshire, A., & Dalgleish, T. (2011). Extending brain-training to the affective domain: Increasing cognitive and affective executive control through emotional working memory training. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024372

Silk, J. S., Davis, S., McMakin, D. L., Dahl, R. E., & Forbes, E. E. (2012). Why do anxious children become depressed teenagers? the role of social evaluative threat and reward processing. *Psychological Medicine*, *42*(10), 2095–2107. https://doi.org/10.1017/s0033291712000207

Sfärlea, A., Buhl, C., Loechner, J., Neumüller, J., Asperud Thomsen, L., Starman, K., Salemink, E., Schulte-Körne, G., & Platt, B. (2020). "I am a total…loser" – the role of interpretation biases in youth depression. *Journal of Abnormal Child Psychology*, 48(10), 1337–1350. https://doi.org/10.1007/s10802-020-00670-3

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.346

Singh, S. (2020). An integrated model combining ECM and UTAUT to explain users' post-adoption behavior towards Mobile Payment Systems. *Australasian Journal of Information Systems*, 24. https://doi.org/10.3127/ajis.v24i0.2695

Smith, L., Leung, W. G., Crane, B., Parkinson, B., Toulopoulou, T., & Yiend, J. (2017). Bilingual comparison of Mandarin and English cognitive bias tasks. *Behavior Research Methods*, *50*(1), 302–312. https://doi.org/10.3758/s13428-017-0871-0

Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology*, 44(5), 1075–1082. https://doi.org/10.1037//0022-3514.44.5.1075

Stade, E. & Stirman, S. & Ungar, L. & Schwartz, H. & Yaden, D. & Sedoc, J. & DeRubeis, R. & Willer, R. & Eichstaedt, j. (2023). Artificial intelligence will change the future of psychotherapy: A proposal for responsible, *psychologist-led development*. 10.31234/osf.io/cuzvr.

Stafford, R. Q., Broadbent, E., Jayawardena, C., Unger, U., Kuo, I. H., Igic, A., Wong, R., Kerse, N., Watson, C., & MacDonald, B. A. (2010). Improved robot attitudes and emotions at a retirement home after meeting a Robot. *19th International Symposium in Robot and Human Interactive Communication*. https://doi.org/10.1109/roman.2010.5598679

Stahl, S. M. (2002). The psychopharmacology of energy and fatigue. Journal of Clinical Psychiatry, 63(8), 7-8.

Starreveld, P. A., & La Heij, W. (2016). Picture-word interference is a Stroop effect: A theoretical analysis and new empirical findings. *Psychonomic Bulletin & Review*, *24*(3), 721–733. https://doi.org/10.3758/s13423-016-1167-6

Steger, M. F., & Kashdan, T. B. (2009). Depression and everyday social activity, belonging, and well-being. *Journal of Counseling Psychology*, 56(2), 289–300. https://doi.org/10.1037/a0015416

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103. https://doi.org/10.1207/s15327752jpa8001_18

Steinman, S. A., & Teachman, B. A. (2014). Reaching new heights: Comparing interpretation bias modification to exposure therapy for extreme height fear. *Journal of Consulting and Clinical Psychology*, 82(3), 404–417. https://doi.org/10.1037/a0036023

Stroud, C., Walker, L. R., Davis, M., & Irwin, C. E. (2015). Investing in the health and well-being of Young Adults. *Journal of Adolescent Health*, *56*(2), 127–129. https://doi.org/10.1016/j.jadohealth.2014.11.012

Sutton, T. M., Altarriba, J., Gianico, J. L., & Basnight-Brown, D. M. (2007). The automatic access of emotion: Emotional stroop effects in Spanish–English bilingual speakers. *Cognition and Emotion*, 21(5), 1077–1090. https://doi.org/10.1080/02699930601054133

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2011). Language-related ERP components. *Oxford Handbooks Online*. https://doi.org/10.1093/oxfordhb/9780195374148.013.0197

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-Assisted Research Design and analysis*. Allyn and Bacon

Tam, C., Santos, D., & Oliveira, T. (2018). Exploring the influential factors of continuance intention to use mobile apps: Extending the expectation confirmation model. *Information Systems Frontiers*, 22(1), 243–257. https://doi.org/10.1007/s10796-018-9864-5

Tam, N. W., Kwok, S. Y., & Gu, M. (2023). Individual, peer, and family correlates of depressive symptoms among college students in Hong Kong. *International Journal of Environmental Research and Public Health*, *20*(5), 4304. https://doi.org/10.3390/ijerph20054304

Tamilmani, K., Rana, N. P., & Dwivedi, Y. K. (2020). Consumer acceptance and use of information technology: A meta-analytic evaluation of UTAUT2. *Information Systems Frontiers*, 23(4), 987–1005. https://doi.org/10.1007/s10796-020-10007-6

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The New Era of mixed methods. *Journal of Mixed Methods Research*, *1*(1), 3–7. https://doi.org/10.1177/2345678906293042

Tamilmani, K., Rana, N. P., & Dwivedi, Y. K. (2018). Use of 'habit' is not a habit in understanding individual technology adoption: A review of UTAUT2 based empirical studies. *Smart Working, Living and Organising*, 277–294. https://doi.org/10.1007/978-3-030-04315-5_19

Teachman, B. A., Joormann, J., Steinman, S. A., & Gotlib, I. H. (2012). Automaticity in anxiety disorders and major depressive disorder. *Clinical Psychology Review*, *32*(6), 575–603. https://doi.org/10.1016/j.cpr.2012.06.004

The Family Planning Association of Hong Kong. (2023). *The Report of Youth Sexuality Study 2021*. https://www.famplan.org.hk/zh/media-centre/press-releases/detail/report-on-youthsexuality-study-2021-secondary-school-survey

Titov, N., Andrews, G., Johnston, L., Robinson, E., & Spence, J. (2010). Transdiagnostic internet treatment for anxiety disorders: A randomized controlled trial. *Behavior Research and Therapy*, *48*(9), 890–899. https://doi.org/10.1016/j.brat.2010.05.014

Tong, A. C., Tsoi, E. W., & Mak, W. W. (2021). Socioeconomic status, mental health, and workplace determinants among working adults in Hong Kong: A latent class analysis. *International Journal of Environmental Research and Public Health*, *18*(15), 7894. https://doi.org/10.3390/ijerph18157894

Torkan, H., Blackwell, S. E., Holmes, E. A., Kalantari, M., Neshat-Doost, H. T., Maroufi, M., & Talebi, H. (2014). Positive imagery cognitive bias modification in treatment-seeking patients with major depression in Iran: A pilot study. *Cognitive Therapy and Research*, *38*(2), 132–145. https://doi.org/10.1007/s10608-014-9598-8

Trochim, W.M.K. (2006). *Research methods knowledge base*. Retrieved on April 26, 2024 from http://www.socialresearchmethods.net

Trost, M. J., Chrysilla, G., Gold, J. I., & Matarić, M. (2020). Socially-assistive robots using empathy to reduce pain and distress during peripheral IV placement in children. *Pain Research and Management*, *2020*, 1–7. https://doi.org/10.1155/2020/7935215

Tzavella, L., Maizey, L., Lawrence, A. D., & Chambers, C. D. (2020). The affective priming paradigm as an indirect measure of food attitudes and related choice behavior. *Psychonomic Bulletin & Review*, *27*(6), 1397–1415. https://doi.org/10.3758/s13423-020-01764-1

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in Mental Health: A review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Vaismoradi, M., Turunen, H., & Bondas, T. (2013a). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences*, 15(3), 398–405. https://doi.org/10.1111/nhs.12048

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2013). A gentle introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, 85(3), 842–860. https://doi.org/10.1111/cdev.12169

van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (2022). *Bayesian Repeated-Measures ANOVA: An Updated Methodology Implemented in JASP*. https://doi.org/10.31234/osf.io/fb8zn

van Vugt, H. C., Hoorn, J. F., & Konijn, E. A. (2009). Interactive engagement with Embodied Agents: An empirically validated framework. *Computer Animation and Virtual Worlds*, *20*(2–3), 195–204. https://doi.org/10.1002/cav.312

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of Technology. *MIS Quarterly*, *36*(1), 157. https://doi.org/10.2307/41410412

Wada, K.., & Shibata, T. . (2007). Living with seal robots—its sociopsychological and physiological influences on the elderly at a Care House. *IEEE Transactions on Robotics*, *23*(5), 972–980. https://doi.org/10.1109/tro.2007.906261

Wang, X., Wang, Y., & Xin, T. (2020). The psychometric properties of the Chinese version of the Beck Depression Inventory-II with Middle School Teachers. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.548965

Wang, Z., Chen, J., Boyd, J. E., Zhang, H., Jia, X., Qiu, J., & Xiao, Z. (2011b). Psychometric Properties of the Chinese version of the perceived stress scale in policewomen. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028610

Wang, Z., Yuan, C.-M., Huang, J., Li, Z.-Z., Chen, J., Zhang, H.-Y., Fang, Y.-R., & Xiao, Z.-P. (2011a). Reliability and validity of the Chinese version of Beck Depression Inventory-II among depression patients. *Chinese Mental Health Journal*, 25(6), 476–480.

Wenzlaff, R. M. (1988). How to unmask depressive thinking: The role of attentional focus in depression. *Journal of Abnormal Psychology*, 97(1), 68-74.

Wenzlaff, R. M. (1993). The mental control of depression: Psychological obstacles to emotional well-being. In D. M. Wegner & J. W. Pennebaker (Eds.), *Handbook of mental control* (pp. 239-257). Prentice Hall.

Wenzlaff, R. M., & Bates, D. E. (1998). Unmasking a cognitive vulnerability to depression: how lapses in Mental Control reveal depressive thinking. *Journal of Personality and Social Psychology*, *75*(6), 1559–1571. https://doi.org/10.1037//0022-3514.75.6.1559

WHO. (2023). *Depressive disorder (depression)*. https://www.who.int/news-room/fact-sheets/detail/depression.

Williams, G. J. M., Watts, F. N., MacLeod, C., & Mathews, A. (1988). *Cognitive psychology and emotional disorders*. John Wiley & Sons.

Wisco, B. E. (2009). Depressive cognition: Self-reference and depth of processing. *Clinical Psychology Review*, *29*(4), 382–392. https://doi.org/10.1016/j.cpr.2009.03.003

Wisco, B. E., & Nolen-Hoeksema, S. (2010). Interpretation bias and depressive symptoms: The role of self-relevance. *Behavior Research and Therapy*, *48*(11), 1113–1122. https://doi.org/10.1016/j.brat.2010.08.004

Wood, D., Crapnell, T., Lau, L., Bennett, A., Lotstein, D., Ferris, M., & Kuo, A. (2017). Emerging adulthood as a critical stage in the life course. *Handbook of Life Course Health Development*, 123–143. https://doi.org/10.1007/978-3-319-47143-3_7

Würtz, F., Zahler, L., Blackwell, S. E., Margraf, J., Bagheri, M., & Woud, M. L. (2022). Scrambled but valid? the scrambled sentences task as a measure of interpretation biases in psychopathology: A systematic review and meta-analysis. *Clinical Psychology Review*, *93*, 102133. https://doi.org/10.1016/j.cpr.2022.102133

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2909824.3020230

Yeh, C. (2004). The Relationship of Cohesion and Coherence: A Contrastive Study of English and Chinese.

Yi, Y. (1990). A Critical Review of Consumer Satisfaction. In V. A. Zeithaml (Ed.), Review of Marketing 1990 (pp. 68-123). Chicago, IL: American Marketing Association.

Yiend, J., André, J., Smith, L., Chen, L. H., Toulopoulou, T., Chen, E., Sham, P., & Parkinson, B. (2019). Biased cognition in East Asian and Western cultures. *PLOS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0223358

Yiend, J., Lee, J.-S., Tekes, S., Atkins, L., Mathews, A., Vrinten, M., Ferragamo, C., & Shergill, S. (2013). Modifying interpretation in a clinically depressed sample using 'cognitive bias modification-errors': A double blind randomised controlled trial. *Cognitive Therapy and Research*, *38*(2), 146–159. https://doi.org/10.1007/s10608-013-9571-y

Yuan, S., Ma, W., Kanthawala, S., & Peng, W. (2015). Keep using my health apps: Discover users' perception of health and fitness apps with the UTAUT2 model. *Telemedicine and E-Health*, *21*(9), 735–741. https://doi.org/10.1089/tmj.2014.0148

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d18-1009

Zhang, D., Shen, J., Li, S., Gao, K., & Gu, R. (2021). I, robot: Depression plays different roles in human–human and human–robot interactions. *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/s41398-021-01567-5

Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S., & Lan, P. (2017). Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. *Annals of translational medicine*, 5(4).

Zhang, S. and Ahn, C. 2013. Sample Size Calculations for Comparing Time-Averaged Responses in K-Group Repeated-Measurement Studies. *Computational Statistics and Data Analysis*. Vol 58(1). Pages 283-291.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1–38. https://doi.org/10.1145/3639372

Zhou, T., Lu, Y., & Wang, B. (2010). Integrating TTF and UTAUT to explain mobile banking user adoption. *Computers in Human Behavior*, 26(4), 760–767. https://doi.org/10.1016/j.chb.2010.01.013

Zhu, J., Zhang, J., Sheng, Z., & Wang, F. (2018). Reliability and validity of the Beck Depression Inventory-II applied to Chinese construction workers. *Social Behavior and Personality: An International Journal*, *46*(2), 249–258. https://doi.org/10.2224/sbp.6638

Zhu, S., Tse, S., Goodyear-Smith, F., Yuen, W., & Wong, P. W. (2016). Health-related behaviors and mental health in Hong Kong employees. *Occupational Medicine*, *67*(1), 26–32. https://doi.org/10.1093/occmed/kqw137