Investigating User Experience of Chatbot Repair Strategies in Simple Versus Complex Tasks

by

Aaron Rios

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2022 by the
Graduate Supervisory Committee:

Nancy J. Cooke, Chair
Erin K. Chiou
Robert S. Gutzwiller

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

The implementation of chatbots in customer service is widely prevalent in today's world with insufficient research to appropriately refine all of their conversational abilities. Chatbots are favored for their ability to handle simple and typical requests made by users, but chatbots have proven to be prone to conversational breakdowns. The study researched how the use of repair strategies to combat conversational breakdowns in a simple versus complex task setting affected user experience. Thirty participants were collected and organized into six different groups in a two by three between subjects factorial design. Participants were assigned one of two tasks (simple or complex) and one of three repair strategies (repeat, confirmation, or options). A Wizard-of-Oz approach was used to simulate a chatbot that participants interacted with to complete a task in a hypothetical setting. Participants completed the task with this researcher-controlled chatbot as it intentionally failed the conversation multiple times, only to repair it with a repair strategy. Participants recorded their user experience regarding the chatbot afterwards. An Analysis of Covariance statistical test was run with task duration being a covariate variable. Findings indicate that the simple task difficulty was significant in improving the user experience that participants recorded whereas the particular repair strategy had no effect on the user experience. This indicates that simpler tasks lead to improved positive user experience and the more time that is spent on a task, the less positive the user experience. Overall, results associated with the effects of task difficulty and repair strategies on user experience were only partially consistent with previous literature.

DEDICATION

I want to recognize the support and optimism that my family has always displayed in my academic goals since I was a child. They instilled in me the importance of prioritizing my education and I have never stopped to question the value that I gain with every lesson that I encounter. Thank you all for being by my side in every step of the way.

ACKNOWLEDGMENTS

I want to thank my thesis committee who have supported and guided me in this thesis. Not only have they been constructive and insightful, but they were consistently willing to offer ideas and feedback throughout the entirety of this process. I can only regret not getting the chance to further learn from them during such a busy time, and I can only hope for further opportunities to collaborate at a deeper level with each of them in the future! They are Dr. Nancy Cooke, Dr. Erin Chiou, and Dr. Robert Gutzwiller. Thank you to each of you for your expertise and care.

I would like to offer a special thank you to my chair, Dr. Cooke, for not only the mentorship she has provided me, but also the unique opportunities that she has introduced to me. She has greatly contributed to my academic and professional growth within the Human Systems Engineering program. I can attribute this to her willingness to extend an invitation to me to share in her interests as a member of her Center for Human, Artificial Intelligence, and Robot Teaming (CHART) Lab. Thank you Dr. Cooke!

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Background

The first chatbot, Eliza, was developed in the 1960s (Mitchell & Mamykina, 2021). In that time, a chatbot, such as Eliza, could simply return user inputs as questions and perform low-level pattern matching (Bradeško & Mladenić, 2012). The conversational ability was unique for its time, but still very poor. Since then, chatbots have enhanced their conversational capabilities by use of various forms of data, artificial intelligence, and machine learning. However, this level of technology is not present in every chatbot and the inadequate conversational capabilities of chatbots are still prevalent. Regardless, chatbots that are being broadly implemented into a vast number of domains in today's world. One domain that popularly utilizes these chatbots is customer service (Li et al., 2019), and this domain is expected to be the primary beneficiary of chatbot technology according to Duijst (2017). Rapp (2021) predicted that 85% of the customer service interactions in 2022 would occur with a customer service chatbot. It is worth noting that users' self-reporting of the rationale behind usage of chatbots for customer service has been most highly associated with the user aiming for a quick and efficient response for a narrowed question online. Additionally, Rapp (2021) noted that acceptance of chatbots for functional purposes is dependent on the skills that are highly valued within a particular field. Researchers who have spent time investigating these chatbots in customer service have therefore put an emphasis on the way that people are able to successfully complete their tasks in a timely manner.

With the massive role that chatbots are rapidly acquiring in customer service it is vital that research be done to improve their conversational ability and user experience. Furthermore, the level of in-depth research in this area has not kept pace with this rapid trend (Følstad & Skjuve, 2019). One way to contribute to chatbots at all levels of maturity might be to understand how they could overcome failures in their own conversational breakdowns. Understanding and implementing strategies that can repair conversational failures effectively within chatbots is one realistic solution to further control the negative experiences of poor interaction.

Importance of Problem

The use of chatbots, a type of conversational agent (CA), in task-oriented customer service settings has grown tremendously in recent years and has only shown signs of increased utilization in the future (Li et al., 2020a). Conversational agents (CA) are generally used by organizations and companies to assist a user in completing a task while minimizing the need for a human worker in some or all of the process (Lee & Choi, 2017). These automated conversational agents (CA) interact with a user to fulfill two responsibilities: 1) demonstrate competent understanding of the user's inquiry or request, 2) fulfill that user's inquiry or request (Chattaraman et al., 2019). However, common problems among conversational agents, both verbal and text-based, typically involve the CA failing to interpret the literal terms of the request being made (a speech recognition issue) and/or failing to interpret the meaning of the request in the given context of the interaction (Jung et al., 2020). Such failures in task-oriented settings, in which communication is essential to goal attainment, tend to lead to the conversational agent's inability to fulfill user requests, thereby rendering them ineffective (Ashktorab et al.,

2019). An ineffective interaction has been shown to lower user experience and increase the unwillingness to disclose personal information to the CA (Guo et al., 2021). In addition, when faced with a non-progressive exchange with a chatbot the likeliest measured outcome of user behavior was to quit the topic at hand (i.e., asking a new question, leaving the conversation). Users' tendency to completely abandon the interaction is significantly likelier after just one of these non-progressive exchanges (Li et al., 2019; Li et al., 2020b). For this reason, it is important that CAs are equipped with repair strategies that improve the capabilities of the CA to better process and complete user inquiries (Guo et al., 2021).

Repair strategies are strategies used to repair conversational breakdowns (also known as non-progressive exchanges) to allow for the completion of the intended goal in an interaction (Ashktorab et al., 2019). Conversational breakdowns occur when a conversational agent fails to interpret a user's intended meaning in an exchange (Li et al., 2020a). This can also occur on the user-end of the interaction if the user chooses not to respond to the agent for any given exchange. Conversational breakdowns can suspend the conversation until further clarification or rephrasing is done in a manner that the agent can understand. Most conversational agents already have repair strategies in place to improve the interaction if the agent cannot process the users' requests (Ashktorab et al., 2019). Some instances of this include a chatbot asking the user to reword their response or having the user choose from a list of potential interpretations that the chatbot generated. However, it is of value to understand how these repair strategies are perceived by users. Despite the common use of repair strategies in chatbots today, there is an insufficient amount of research that focuses on the ramifications of their use (Guo et al.,

2021). In this way, the findings may elicit justifiable use of specific repair strategies in cases where a chatbot is not advanced enough to solve user requests immediately. It may also offer insight on the impact of reusing repair strategies, consecutively, following multiple conversational breakdowns.

Research Questions

These research questions were addressed: 1) *How does the type of repair strategy used by a chatbot during a conversational breakdown affect user experience*? 2) *How does the difficulty of a task affect user experience across various repair strategies?*

CHAPTER 2

LITERATURE REVIEW

Introduction

The research surrounding chatbot interfaces and conversational agents (CAs) is abundant, but the still limited capabilities of conversational agents suggest that there are gaps in the understanding of efficient chatbot design. Two predominant gaps needing to be addressed are the methodology of implementing repair strategies in chatbots and the effectiveness of the varying strategies in real-world scenarios (Guo et al., 2021). The reviewed literature for the proposed study focuses on the following three our topics: 1) chatbot interfaces, 2) user experience with chatbots, and 3) task difficulty.

Chatbots

Background

Chatbots are classified as text-based conversational agents. They primarily communicate with users through text-based messages on a messaging platform. Facebook Messenger, Kik, Skype, and many other platforms can host the interaction between a human and a chatbot. With the variety of platforms and domains that chatbots are implemented in there comes a variety of requests that chatbots are developed to fulfill.

Task-oriented versus Social-oriented

Chattaraman and colleagues (2019) define two interaction styles that CAs can model to better fulfill these varying user requests. Depending on the goals of the user, a chatbot should look to implement either a social-oriented or task-oriented interaction style. A CA that utilizes a social-oriented interaction style tends to use informal dialog and often includes social topics that are not related to the task to meet socioemotional

goals. This can be in the form of small talk, emotional conversation, or other positive expressions. A CA that utilizes a task-oriented interaction style tends to be more formal and conversation is often focused only on topics that help achieve the task at hand. Chattaraman, et al. (2019) observed a virtual assistant (both verbal and text-based) that interacted with older individuals. Though with virtual assistants, the results appear generalizable to chatbot CAs and suggest that as the difficulty of user requests increases the more that a task-oriented chatbot is desired. The simpler or less demanding that a task is the more likely that users are willing to interact with a social-oriented CA. Given the task-driven context of the current study a task-oriented chatbot was designed to interact with participants and thus chatbots that were task-oriented were the primary focus of the reviewed literature.

Conversational Breakdowns and Repair Strategies

Atefi and Alipour (2019) conducted research on automated testing frameworks for conversational agents. Their work recognized three categories of conversational breakdowns that chatbots may have in an interaction with a user. The first type of conversational breakdown occurs when a CA misunderstands the user's request. In this case a chatbot may interpret an inaccurate concept or question and offer a completely irrelevant response back. The second type of breakdown is when a CA cannot answer a user's request. The chatbot may understand the context or even the question being asked but cannot put an answer into words. A chatbot response to a user may involve sending relevant links that can be associated with the recognized keywords from the user's request. The final breakdown occurs when a CA's intention or purpose is not interpretable by the user. Atefi and Aipour give an example of this breakdown where a

user asks a CA, “Do you know that movie will be aired on Friday night?” and the CA responds by saying, “Yes, yes,” In this case the users have received a response that is not clearly off topic, but the users also have no way of knowing if the CA knows about the movie airing on Friday night. The CA did not offer any information about the movie and while the CA claims that it knows the answer to their question, it offers a response that does not confirm that this is the case. These are some of the types of conversational breakdowns that are addressed by the use of repair strategies.

Li and colleagues (2021) aimed to specifically study issues with intelligent agents that occur because of natural language understanding errors - specifically the categories of intent detection and slot value extraction within them. Here, the former is defined by a misunderstanding of the idea that the user is attempting to articulate. The latter, relates to the way that agents do not accurately respond to the user’s request. These are notions described by Jung and colleagues (2021) also as their research highlighted three main problems that users had with CA behavior. The first is that CAs were too repetitive. The other two are that the CA cannot understand the exact meaning of what they say and that the CA cannot recognize what was actually being said. In this way, we see a clear necessity for the current research.

Eight Types of Repair Strategies

Although sufficient research relevant to repair strategies in chatbots has not yet been conducted, there is previous and relevant information that is applicable to chatbots. Ashktorab and colleagues (2019) identify eight repair strategies that could be used in chatbot design. These strategies are described below:

Top Response. This the only response described that does not acknowledge that a conversational breakdown has occurred, but instead gives the top recommendation based on the system's rules for what response is associated with the highest confidence.

Repeat. This is a response that acknowledges that a conversational breakdown has occurred and simply repeats the prompt it previously stated before the breakdown.

Confirmation. This is a response that acknowledges that a conversational breakdown has occurred and looks to confirm what the understood utterance/request was with the user.

Options. This is a response that acknowledges that a conversational breakdown has occurred and offers the user options that state the most likely utterances/requests that the user made.

Defer. This is a response that acknowledges that a conversational breakdown has occurred and simply transfers the user to a human representative to solve the issue.

Keyword Highlight Explanation. This is a response that acknowledges that a conversational breakdown has occurred and highlights the keywords that the system is focusing on and understanding. This allows users to understand what the CA is interpreting, and they can rephrase their request to remove the misunderstood or confusing words.

Keyword Confirmation Explanation. This is a response that acknowledges that a conversational breakdown has occurred and has the same processes as the *Keyword Highlight Explanation* strategy. The only difference is that the CA will state the understood keywords and ask if the user can confirm its accuracy. This strategy adds an additional exchange to the conversation.

Out-of-vocabulary Explanation. This is a response that acknowledges that a conversational breakdown has occurred and highlights words that the system cannot understand or recognize so that the user can rephrase their request without using them.

Upon identifying these eight repair strategies Ashktorab and colleagues (2019) observed 203 participants in a pairwise comparison experiment. Participants were shown multiple sets of two scenario-based interactions between a user and a chatbot. The chatbots in each pair implemented two different repair strategies and participants were asked to choose which interaction they preferred. They found that participants preferred interactions where chatbots offered users multiple choices and explanations about how to address conversational breakdowns. They also identified aspects of the studied repair strategies that users found to be effective. These included: efficiency and efficacy, intelligence and capability, politeness, and naturalness. Of the specific repair strategies attempted by the chatbots, participants preferred the *options* strategy most in both successful and unsuccessful (failing to achieve the specified goal) interactions. For unsuccessful interactions, participants next preferred the *defer* repair strategy. Therefore, the researchers concluded that though users are typically initially willing to resolve the issue with the agent, they often prefer a human if the attempt is not successful. Such information directly relates to the way that useful repair strategies may ultimately make users more willing to interact with agents overall - as described by Guo and colleagues (2021).

Rule-Based Interaction

Chatbots tend to be built to approach interaction in a rule-based manner or a data-driven one. A rule-based approach involves a chatbot responding to a request based on

prefixed rules. This approach relies on identifying preset rules (such as identifying a key word) and applying a transform rule that will allow for unique responses given the patterns associated with that identified rule. For example, if a CA identifies a word in a request that is considered an emotion the transform rule will bring about responses that are associated with the identified emotion (the pattern that the CA's would recognize and follow). This allows for a response that is relevant to the user's request. These patterns and rules are developed for CA's by using formal logic rules. These logic rules act as proofs that form a conclusion with the goal of proving or disproving a particular understanding based on deductive reasoning. In this manner, user requests can be observed for key items and the logical rules can use predetermined logic to decide what transform rule to apply based on that identification. From there a response or pattern of behavior from the CA can be applied to the interaction. This approach requires handcrafted rules and could potentially take an extensive amount of time to develop for a chatbot that needs to handle complex tasks (Mnasri, 2019). This is the approach used to develop the chatbot within the current study.

Mnasri also describes what a data-driven approach would mean for chatbots' decisions about their behaviors/responses. This approach is more technology based and uses the data from existing documents or previous conversation to create chatbot responses. This approach requires a method for the chatbot to practice information retrieval, machine learning, or a mixture of both capabilities to effectively develop responses.

User Experience

There are several methods for how user experience is measured and how these different components relate to each other. User experience was determined in the current study by observing just three of the several components that can be tied to user experience. These three components are usability, perceived usefulness, and satisfaction. The application and measure of these components in regard to chatbots is the focus of the literature in this section.

Usability

It is important to establish that usability is a broad concept that is applicable to many areas. Usability and user experience are researched quite frequently in the field of human-computer interaction (HCI). Dillion (2001) defines usability in his research as, "the effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in particular environments." There are several tools and models that have been developed, tested, and implemented to better understand how users interact with technology. For example, Davis' (1989) technology acceptance model aimed to use self-reporting measures of users' perceived usefulness and perceived ease of use as means to understand the way that one intends to use a piece of technology. Since the original development of the technology acceptance model, there have been several attempts to better understand the way that subjective measures can be used to understand how users feel toward using a particular piece of technology (Partala & Saari, 2015; Venkatesh & Bala, 2008; Venkatesh et al., 2003). Usability has been one of the metrics that continuously is used as an evaluative tool in these settings and with chatbots in HCI.

Duijst (2017) also found usability to be a key evaluative tool to measure the effect of personalization and task difficulty on user experience.

Perceived Usefulness

Partala and Saari (2015) studied the way the emotions (particularly, valence) tie into the way that users tend to feel about and adopt technology. The researchers found that one major indictor of successful technology adoption was the degree to which users perceive the technology to be useful to them specifically. In such cases, the researchers emphasize that the absence of negative emotions, particularly frustration, is key to successful adoption and that users feeling negatively toward the technology directly leads to their perception of poor user experience with a product or service. Regardless, of the measure of user experience, the product being measured, or the service that is being evaluated, perceived usefulness is fundamental to the way people perceive technology. Perceived usefulness is one of the three primary focuses of the User Experience Instrument used in the current study.

Satisfaction

This research is fundamental to the study at hand. One area of literature surrounding user experience and chatbots has focused on the way that increased frustration leads to decreased satisfaction and therefore decreased user experience. Chaves' (2020) review of chatbot literature found that strong, and enjoyable, human-chatbot interactions intentionally aim to avoid negative emotions. Avoiding these negative emotions becomes possible through a consistent demonstration of flowing conversation with limited conversational breakdowns. Chaves noted that users expect chatbots to be capable of navigating through the conversation in a smooth way and can

feel considerable disappointment if they perceive the chatbot to be underperforming. In application, conscientiousness is noted as being able to comprehend a users' utterances as well as to avoid conversational breakdowns. A decreased sense of conscientiousness is therefore associated with negative emotions, lowered satisfaction, and therefore lower user experience overall.

Task Difficulty

According to the findings by Chaves and Gerosa's (2020) study, increased task difficulty can be linked to decreased levels of user experience. They recognized research by Dyke and colleagues (2013), as well as Duijst (2017), that found that usability and satisfaction tend to be higher when tasks are shorter and simpler. This is because these tasks usually involve less mistakes and more possibility for simple mistake corrections.

Outside of chatbots, with more focus on a task alone, Tavakoli (2009) describes that the more information provided for a task, the more complex it is perceived. Additionally, research done by Goldhammer and colleagues (2014) conclude that time is a function of task difficulty.

Hypotheses

According to the findings of Ashktorab's and colleagues (2019) the *options* repair strategy was the most preferred repair strategy of the eight defined strategies in each of the conditions regardless of a successful or unsuccessful interaction. The participants' demonstration of similar preference for the *options* repair strategy in the current study's live interaction (versus other strategies) would suggest that they would be more likely to indicate a positive user experience (Cockburn, et al., 2017). Cockburn and colleagues directly tie subjective experiences, such as preference, to increased satisfaction as well as

other components of positive user experience that are not directly measured in the current study. With this in mind the hypothesis for RQ1 is:

Hypothesis (H1): Participants that receive the *Options* repair strategy will report better user experience than those who receive the *Repeat* or *Confirmation* repair strategies because it promotes efficiency and freedom of choice (Ashktorab et al., 2019).

Duijst (2017) partially attributes an increase in positive user experience in the simple task condition to the shorter duration of the task. This led to less mistakes and easier correction of any mistakes that did occur and in turn, was the reason for increased usability and less frustration, thereby increasing satisfaction. Chaves and Gerosa (2020) also find support for this claim in their literature review. Therefore, the hypothesis for RQ2 is:

Hypothesis (H2): Participants that receive a simple task will report better user experience than those who were assigned a complex task because a simple task will improve usability and satisfaction. (Duijst, 2017).

CHAPTER 3

METHODOLOGY

Design

The current study implemented a 2x3 between subjects factorial design with a posttest. The treatments were implemented to investigate the impact of the type of repair strategy on user experience given conversational breakdowns. The repair strategy type and task difficulty were the treatments in this study. Each study session implemented the treatments by assigning one of three repair strategies and either a simple or complex task to a participant. A researcher-controlled "chatbot" interacted with participants while inducing scripted conversational breakdowns so that repair strategies could be implemented to fix them. After each task, participants' user experience was measured with a subjective, Likert-based instrument. A between-subjects analysis of covariance (ANCOVA) was the statistical test used to analyze the data collected while using task duration as a covariate variable.

Participants

An a priori power analysis was conducted using G*Power3 (Faul et al., 2007) to test the difference between the means of 3-repair strategies by 2-task difficulty using an *F-test*, with a large effect size ($\eta_p^2 = 0.14$; Cohen, 1988), and an alpha (α) of 0.1. According to the result, 70 participants are needed to run the experiment. However, only 30 participants were recruited for this study. See Appendix A for the details of the analysis.

The participants recruited were students from Arizona State University who were associated with the Arizona State University SONA pool. All participants were at least

eighteen years of age. This sample consisted of twenty-nine participants between the age of eighteen to twenty-four and one participant between the age of thirty-five to forty-four.

Variables

Independent Variables

The first independent variable was the type of repair strategy used to address a conversational breakdown that occurs between the participant and the Bluewidth Internet chatbot. These repair strategies were implemented an equal number of times, once per study, among the simple and complex task conditions. The three types of repair strategies being tested were: *repeat*, *confirmation*, and *options*.

These three repair strategies were the focus of the current research for two major reasons. The first reason is that they acknowledge a conversational breakdown with a user. This circumstance is necessary to highlight a non-progressive exchange to allow for the implementation of a clear repair strategy that the user can recognize. This would allow for better control and measure of the effect of the repair strategy in the users' responses. The second reason is that they are the most practical to test of the ones acknowledging breakdowns. Of the eight repair strategies outlined by Ashktorab and colleagues (2019), the *top response* strategy is the only one that does not acknowledge when a conversational breakdown occurs in an interaction with a user, therefore, it was not tested in the current study. The *defer* strategy would immediately end the interaction after one conversational breakdown and in turn, would not allow for the measure of the dependent variable so it was excluded from the current study. The *keyword highlight explanation*, *keyword confirmation explanation*, and *out-of-vocabulary explanation* strategies were not testable given the lack of control that would be introduced given the

current study design where a Wizard-of-Oz approach is used to replicate chatbot technology. This left the three chosen repair strategies that were comparable and immediately applicable to commonplace chatbots that contain less advanced capabilities.

The second independent variable was task difficulty. Task difficulty is either defined by being simple or complex. Participants were randomly assigned either a simple task or complex task at the beginning of the study.

The tasks were classified as simple or complex based on the amount of information provided for the task and the time of task completion (task duration). The design of the tasks consider the reviewed literature provided about task difficulty from both Tavakoli (2009) and Goldhammer and colleagues (2014). The complex task has over twice the amount of lines of contextual information than the simple task. This would imply that the perceived task difficulty is greater in the complex task than the simple task (Tavakoli, 2009). The complex task also has at least twice the amount of required steps to reach task completion than the simple task; this would, given this logic, mean that the complex task would require more time for completion than the simple task. Therefore, implying that the complex task is more difficult than the simple task (Goldhammer et al., 2014).

The decision to test the simple versus complex task was done to further observe its effect on user experience given Duijst's (2017) claim of significance while also attempting to connect repair strategies to these two constructs. This research was fundamental to the current study because it observed two desired components, user experience and task difficulty. The similar design to the current research allowed for the substitution of personalization (one of their independent variables) with the repair

strategy conditions to further validate the instrument with potential for connecting the effects of repair strategy and task difficulty.

Dependent Variable

The dependent variable was user experience. This was measured by observing the three components of user experience with a conversational agent: perceived usefulness, usability, and satisfaction. Previous literature emphasizes the importance of identifying these components to determine participants' user experience within an interaction with a chatbot. These components were measured using a Likert-based questionnaire that is described in this research as the User Experience Instrument (Duijist, 2017).

Demographic Section. This section was provided to the participants at the very beginning of the study. It contained the consent form for the study and 6 demographic questions derived from the work of Duijst (2017) These questions captured the participants' age, level of education, frequency of computer use, confidence in using computers, experience with chatbots, and feelings towards chatbots. See Appendix B for the demographic section.

User Experience Instrument

This section of the questionnaire was provided to participants as a posttest (the dependent variable) at the conclusion of the task/interaction with the chatbot. This section was populated with the metrics measuring participants' user experience given their interaction with the chatbot.

A validated instrument from Duijst (2017) that measured user experience within chatbots was used to observe the effect of the repair strategy and task difficulty conditions. This instrument was constructed from several sources to measure the user

experience metrics: usability, perceived usefulness, and satisfaction (Davis, 1985; Duijst, 2017; Finstad, 2010; McTear et al., 2016) See Appendix C for the instrument.

Materials

Bluewidth Internet Company

Bluewidth Internet is a fictional internet service providing company that was used to create a task-oriented setting. Participants were hypothetical customers of this company and their role consisted of speaking to an automated customer service representative (a chatbot) of Bluewidth Internet. Participants were told to complete their internet-based tasks in this setting with the Bluewidth Internet chatbot to allow researchers to gather information to improve the customer service experience of this "up and coming" company.

Computer

A computer was used to connect the participant with the researcher in the current study using Zoom software. The computer was used to present the participants with the consent form, questionnaire, and the chatbot that they interacted with in a Zoom meeting.

Zoom

Zoom, the video conference software, was used to connect the participant with the experimenter and the chatbot. It recorded the interaction for verbal data and message-based chat data that was received from the exchanges between the participant and the experimenter. The use of the Zoom chat allowed for the researcher to act as an autonomous chatbot in anonymity.

Qualtrics

The survey software, Qualtrics, was used to develop and distribute the questionnaire for the current study. The questionnaire within this software contained the consent form, demographics section and the User Experience Instrument. Participants were directed here to complete it during the study.

Rule-Based Agent Scripts

The conversational agent (chatbot) that the researcher simulated had a script for each repair strategy and for both task difficulties. There was a total of six scripts that were used for six different combinations of task difficulty and repair strategy. One script was needed per study and contained a strict assortment of responses that the chatbot was able to send to the participant during a live study.

The conversation flow was structured to promote rule-based interaction (Singh et al., 2019) and turn-taking protocol as suggested by Toxtli and colleagues (2018). The developed script was also modelled to display level one maturity for a chatbot defined by Duijst's (2017) work. This maturity level indicates that the chatbot primarily follows a simple question and answer conversation flow in a singular manner of communication, being a one-to-one interaction in one language, that is rule-based by word recognition. The chatbot was intentionally designed to be as objectively task-oriented as possible. This is to isolate observable changes in user experience between conditions without introducing user preferences in perceived personality, anthropomorphic traits or other chatbot characteristics as possible confounding variables. The simplistic language of the chatbot was modelled after Ashktorab and colleagues' (2019) work.

In the simple task condition the scripts intentionally induced two conversational breakdowns for the five minimum responses that the participant would need to have to

give to complete the task. In the complex task condition the scripts intentionally induced four conversational breakdowns for the ten minimum responses that the participant would need to complete the task. Based on which repair strategy was assigned to a participant, either the *repeat, confirmation, or options* repair strategy was used to repair all of the conversational breakdowns caused by the chatbot for that one interaction. See Appendix D for these six scripts and see Appendix E for an example diagram of the conversational flow.

Participant Task Information Document

Participants received an online PDF document that provided them with an introduction to the study and a description of their task (either the simple or complex task). This document informed participants of the hypothetical setting in which they have obtained this task, the initial request needed to be made to the Bluewidth Internet chatbot in order to initiate the conversation, and additional information about their task that may assist them in forming requests to the Bluewidth Internet chatbot. Both of the tasks were meant to represent realistic requests that customers might have to make to their internet service provider.

The initial request that participants were instructed to use to start the conversation was enforced as a method for further control in the study. By doing this the Bluewidth Internet chatbot was provided with enough information to invoke conversational breakdowns as well as successful interactions for the entirety of the foreseeable interaction. This allowed the participants to respond as they wanted after this initial request, with the chatbot still being able to advance/breakdown the conversation as

planned with the initial information provided. See Appendix F and Appendix G for the two task information documents that were assigned to the participants.

Procedure

A pilot study was conducted prior to the current study. Six participants were used to estimate the expected task completion time, number of user-end conversational breakdowns, and task failures prior to the current study. Participants first signed up to participate in the proposed study through the ASU SONA system. To participate in the study, the participants entered a Zoom meeting at a scheduled time. The experimenter in the Zoom meeting used the Zoom chat to send them the Qualtrics link to complete the demographic section of the questionnaire. The participant then returned to the Zoom window where the experimenter sent them their task information document and introduced them to the details of the study. The participants were given 2 minutes to read over their task information document. After two minutes, the participants were instructed to message the chatbot (another Zoom profile controlled by the experimenter in the Zoom meeting) to begin the task. The participant interacted with the chatbot, which followed the rule-based agent scripts to guide the interaction, to complete the task in an unspecified amount of time. After the conclusion of the task the participants returned to the questionnaire to complete the posttest. Finally, a debrief was conducted to inform the participant about their performance in completing the assigned task successfully.

CHAPTER 4

RESULTS

There was a total of two data points removed from the collected data set. One participant was removed from the data set for completing the entirety of the questionnaire, including the posttest, in under a minute before receiving the treatment. The second participant that was removed from the data set was the only participant to not reach the conclusion of the task. The participant's failure to carefully read the instructions led to their inability to communicate with the chatbot in a manner that could allow for continuous collaboration. The participant's highly skewed responses reflect a negative user experience that may not be attributed the independent variables, but rather the participant's demonstrated inability to identify the required information to complete the assigned task.

Upon initially identifying a statistically nonsignificant relationship between the independent variables and the dependent variable in an ANCOVA further analysis explored the presence of covariate variables. In this case, task duration, participant task success, and participant task score were tested, however only task duration demonstrated statistical significance. Because of the limited sample size, a *p*-value ranging from 0.05 to 0.1 was categorized as significant, values below 0.05 were categorized as significant, and values below .001 were categorized as highly significant (Gelman, 2013).

A two-way ANCOVA analysis was performed to analyze the effect of repair strategy (*repeat*, *confirmation*, and *options*) and task difficulty (simple and complex) on the mean score of participant user experience. Higher mean scores indicated higher positive user experience. The amount of time, in seconds, spent interacting with the

chatbot to complete the task (task duration) was used as the covariate in this test. An ANCOVA requires that the covariate is highly correlated with the dependent variable. Therefore, a Pearson product-moment correlation was run to determine the relationship between user experience and task duration. There was a negative correlation between user experience and the task duration, which was statistically significant ($r(28) = -0.392.$, $p = 0.032$).

The ANCOVA did not show significant interaction effects between repair strategy and task difficulty, $F(2, 23) = 0.372$, $p = 0.693$. The main effects of the repair strategy condition, $F(2, 23) = 0.032$, $p = 0.968$, was not significant whereas the task difficulty condition, $F(1 ,23) = 3.473$, $p = 0.075$, was considered significant at the $p<0.10$ level. The findings indicate a failure to reject H2 while rejecting H1. See Table 1 for the descriptive statistics.

Table 1

User Experience Score Across Experimental Groups

Task Difficulty	Repair Strategy	Mean	Standard Deviation	n
Simple	Confirmation	4.41	1.74	5
	Options	4.53	1.56	4
	Repeat	4.69	0.21	5
	Total	4.54	1.24	14
Complex	Confirmation	3.74	1.14	5
	Options	4.41	0.85	6
	Repeat	4.09	0.62	5
	Total	4.10	0.88	16
Total	Confirmation	4.07	1.43	10
	Options	4.46	1.10	10
	Repeat	4.39	0.54	10
	Total	4.30	1.07	30

Table 2

User-End Breakdowns by Condition

Task Difficulty	Repair Strategy	Number of Breakdown	Mean (total)	n
Simple	Confirmation	3	0.60	2
	Options	4	1.00	3
	Repeat	3	0.60	1
	Total	10	0.71	6
Complex	Confirmation	8	1.60	3
	Options	7	1.67	3
	Repeat	4	1.00	2
	Total	19	1.25	8
Total	Confirmation	11	1.10	5
	Options	11	1.10	6
	Repeat	7	0.8	3
	Total	29	1	14

Note. The mean reflects the average number of user-end breakdowns for all participants within in each condition. The n value is the number of participants that had one or more user-end breakdowns for that condition (where the maximum is ten for each repair strategy).

CHAPTER 5

DISCUSSION

The current study observed the effect of three specific repair strategies (*repeat*, *confirmation*, and *options*) as well as task difficulty (simple and complex), on the user experience of participants who sought customer support from a hypothetical internet provider. These effects were observed with consideration of the task-time within each tested condition. Given the research on chatbots in similar settings, it was possible to establish key expectations for these relationships. *H1* suggested that the *options* repair strategy would promote the highest reported user experience from participants. *H2* suggested that participants that received a simple task would be more likely to report a higher user experience from the interaction than those who were assigned a complex task.

There were no significant interaction effects found by the analysis of covariance. This describes a lack of notably varied scores of user experience reported by participants because of the presence of both the repair strategy and the task difficulty conditions. The low sample size could be the most realistic explanation for not finding significant effects here.

The ANCOVA indicated that there was no significance found for the repair strategy condition. The *options* repair strategy did not result in the highest reported user experience, therefore, rejecting *H1*. This was not consistent with the findings of Ashktorab and colleagues (2019). The current findings suggested that the *repeat*, *confirmation*, and *options* repair strategies were not any better or worse at improving the user experience that users felt when interacting with a task-oriented chatbot in these circumstances. The low sample size is a potential explanation for this. One other potential

explanation for these results may be due to a lack of sensitivity of the user experience instrument. While it may be an appropriate measure for user experience of task difficulty, it has not been validated for measuring user experience of repair strategies since being created by Duijst (2017). Additionally, the researcher-developed chatbot script may have potentially failed to represent the *options* repair strategy in a way that maximized its effect on user experience. The *options* repair strategy is meant to provide users with the possible chatbot interpretations of their responses. However, the designed chatbot script for this study, while still doing this, focused less on the meaning of the users' responses and more on the actual recognition of the response being said. For example, this chatbot was giving options of specific home addresses that a user might have said rather than options that could be describing where a user said their home may be. In this case the chatbot wants to recognize an exact set of characters and phrases as a response (regardless of their meaning) rather than trying to understand how a user might be describing the concept of where their home could be found (within a country, state, terrain, etc). Failure in identifying meaning and failure to recognize language are both known causes of conversational breakdowns. It is unresearched as to if the *options* repair strategy is more effective in repairing breakdowns caused by chatbots having failures in understanding response meaning or failures in response recognition. Further research would be required to identify if this is the cause of inconsistent findings with past research done by Ashktorab and colleagues (2019).

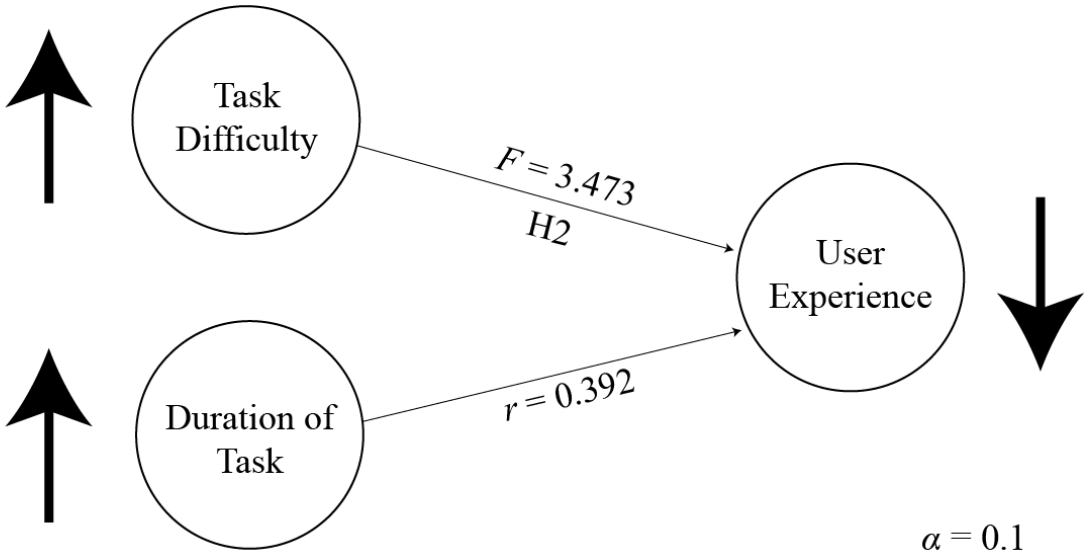
The ANCOVA revealed significance for the task difficulty condition alone. This indicates that the assignment of a simple task or a complex task did significantly alter the level of user experience that participants reported on the user experience instrument. The

analysis supported the finding that the simple task produced better user experience than the complex task. Therefore, it provides support for *H2* and the following claim: the less difficult a task is the greater positive user experience that will be had by users. This claim is consistent with the findings of the literature on task difficulty in chatbots (Chaves & Gerosa, 2020; Duijst, 2017). However, the simpler task also contained less scripted conversational breakdowns than the complex task – even though they occurred in proportion – suggesting that the presence of conversational breakdowns decreases user experience; this is unsurprising, as breakdowns have a proven correlation with negative consequences and decreased user experience (Guo, et al., 2021; Li et al., 2019; Li et al., 2020b).

Given the results of the Pearson product-moment correlation, there is a significant correlation (negative) between the time spent interacting with the chatbot to complete the task (task duration) and the mean user experience score that participants recorded. The relationship found between user experience and task duration can likely be explained by the relationship displayed between task difficulty and task duration. Not only did the simpler task contain fewer conversational breakdowns than the complex task (thereby avoiding increased potential for negative user experience), but task difficulty is negatively (yet significantly) correlated with user experience. Therefore, it is also expected that task duration would have a similar relationship with user experience. Task duration is intentionally designed to be higher in the more difficult task (complex) and lower in the less difficult task (simple). Therefore, as both task duration and task difficulty increase it is implied that the user experience will decrease as a result (negative correlation). See Figure 2 for a visual of this concept.

Figure 2

Shared relationship of task characteristics to UX are consistent with H2



Limitations

The desired sample size was not collected. The power analysis for this study suggested a sample of 70 participants, and therefore was underpowered. The current study only gathered 30 measurable samples. With less than half of the expected sample size collected this limitation has the greatest potential to undermines the validity of the current findings. A potential contributing factor to this shortcoming could be the limited population that was available for participant recruitment as it consisted only of Arizona State University students within a single program.

The lack of generalizability of the findings should be recognized. The participants collected were all students from Arizona State University, each of which were between the ages of 18-24 with the exception of one individual. Further, related research would

need to be conducted on a more diverse set of populations to allow for the generalization of the current findings.

It is necessary to consider the external validity of the findings given the circumstances in which the participants interacted with the chatbot. Mitchell and Mamykina (2021) claimed that most lab settings are not sufficient in understanding user interactions with a chatbot. The occurrence of the current interaction is not long-term nor is in a natural setting to allow for the most realistic user interactions with a customer support chatbot. Short-term research in a lab setting has been shown to produce less accurate user interactions than in an equally observed continuous interaction over time. More realistic observations can be made when there is not an experimenter present to intervene or regulate the interactions. Mitchell and Mamykina (2021) suggest Wizard-of-Oz studies that are more easily conducted in the field over time. While the current study may not have been invalidated by this information, it is important to recognize potential limitations that come with studies that are done in short-term, lab settings such as this one.

Assumptions

There were two assumptions that require recognition given the parameters of the study. The first is that the population being sampled provided their responses with the objective of being honest and attentive. The second is that the time required for a user to respond to the chatbot in nearly every step of the interaction, within both the simple and complex tasks, were relatively similar. This is based on the understanding that all user responses can potentially be answered through copying and pasting responses directly from their task information document. This does not consider the participants' ability to

identify the appropriate information, as all participants were given the same two-minute window to get familiar with the task information. This assumption allows for practical comparisons between the simple and complex task conditions for both task difficulty and task duration.

Future Research

The User Experience Instrument that was used in the current study was derived from the research by Duijst (2017) where significance of the task difficulty (or complexity) was similarly found in its measures of perceived usefulness, usability, and satisfaction. The instrument did not display significance of personalization in Duijst's study and the same held true for the current study when used to measure repair strategies. Future research may explore alternative metrics that consider more and/or different combinations of user experience elements to produce the potential for more effective measure of chatbot repair strategies. Additional research in this area should further test the current results and the findings of Ashktorab and colleagues (2019) to improve the understanding of the effects that repair strategies have when addressing conversational breakdowns**.** The effect of repair strategies on the change in user experience should be tested based on the type of conversational breakdown being address. The presence of user-end breakdowns should also be considered when conducting research on repair strategies despite the lack of a significant role in this study. This should particularly be tested in live interactions between a user and a chatbot as this area is still highly unexplored since the drastic adoption of chatbot technology.

Conclusion

The findings showed that use of the three, promising repair strategies to fix intentional chatbot conversational breakdowns had no significant differential effects on user experience. However, the level of task difficulty that was assigned to participant tasks did present a significant effect in that a simple task was linked to a higher user experience in participants (which correlated with less time they spent on the task and fewer number of breakdowns of the chatbot). The presence of task duration was statistically correlated in a negative manner to the level of task difficulty. These findings do not support the previous research claims about repair strategies in chatbots, but they do align with the research claims made about the use of varying task difficulties. Research about the implementation of repair strategies in chatbot design is limited. The current research could improve the knowledge in chatbot conversational repair to allow for more in-depth research on their use in different task settings to improve user experience. Given the limitations of this study, it is recommended that future research collect more data from a diverse population while also considering the role of the type of conversational breakdown that needs repair.

REFERENCES

Atefi, S., & Alipour, M. (2019). An Automated Testing Framework for Conversational Agents. *Association for Computing Machinery.* https://www.researchgate.net/publication/331195232_An_Automated_Testing_Framework_for_Conversational_Agents/fulltext/5c6b7e674585156b5706b578/An-Automated-Testing-Framework-for-Conversational-Agents.pdf

Ashktorab, Z., Jain, M., Liao, Q., & Weisz, J. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300484

Bradeško, L., & Mladenić, D. (2012). A survey of chatbot systems through a Loebner Prize Competition. *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies.* https://www.researchgate.net/publication/235664166_A_Survey_of_Chatbot_Systems_through_a_Loebner_Prize_Competition

Chattaraman, V., Kwon, W., Gilbert, J., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior (90)* 315–330. https://doi.org/10.1016/j.chb.2018.08.048

Chaves, A., & Gerosa, M. (2020). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, *37*(8), 729–758. https://doi.org/10.1080/10447318.2020.1841438

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Davis, F. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [Unpublished doctoral dissertation]. 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

Dillon, A. (2006). Beyond usability: Process, outcome and affect in human-computer interactions. *The Canadian Journal of Information and Library Science. 26*(4) https://repository.arizona.edu/handle/10150/106391

Dyke, G., Adamson, D., Howley, I., & Rosé, P. (2013). Towards academically productive talk supported by conversational agents. *Intelligent Tutoring Systems*, 531–540. https://doi.org/10.1007/978-3-642-30950-2_69

Duijst, D. (2017). *Can we improve the user experience of chatbots with personalisation?* [Master's thesis, University of Amsterdam]. Research Gate. https://www.researchgate.net/profile/DanielleDuijst/publication/318404775_Can_we_Improve_the_User_Experience_of_Chatbots_with_Personalisation/links/5967ba16a6fdcc18ea662ce7/Can-we-Improve-the-User-Experience-of-Chatbots-with-Personalisation.pdf

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Følstad, A., & Skjuve, M. (2019). Chatbots for Customer Service. *Proceedings of the 1st International Conference on Conversational User Interfaces*. https://doi.org/10.1145/3342775.3342784

Gelman, A. (2013). P values and statistical practice, *Epidemiology*, *11*, 69–72

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

Guo, J., Guo, J., Yang, C., Wu, Y., & Sun, L. (2021). Shing: A conversational agent to alert customers of suspected online-payment fraud with empathetical communication skills. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445129

Jung, H., Kim, H., & Ha, J. (2020). Understanding differences between heavy users and light users in difficulties with voice user interfaces. *Proceedings of the 2nd Conference on Conversational User Interfaces*. https://doi.org/10.1145/3405755.3406170

Lee, S. Y., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, *103*, 95–105. https://doi.org/10.1016/j.ijhcs.2017.02.005

Li, C., Chen, K., & Chang, Y. (2019). When there is no progress with a task-oriented chatbot. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, *59*, 1–6. https://doi.org/10.1145/3338286.3344407

Li, T., Chen, J., Xia, H., Mitchell, T., & Myers, B. (2020a). Multi-modal repairs of conversational breakdowns in task-oriented dialogs. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3379337.3415820

Li, C., Yeh, S., Chang, T., Tsai, M., Chen, K., & Chang, Y. (2020b). A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376209

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface*. Springer.

Mitchell, E., & Mamykina, L. (2021). From the curtain to Kansas: Conducting wizard-of-oz studies in the wild. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411763.3443446

Partala, T., & Saari, T. (2015). Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Computers in Human Behavior*, *53*, 381–395. https://doi.org/10.1016/j.chb.2015.07.012

Singh, J., Joesph, M. H., & Jabbar, K. B. (2019). Rule-based Chabot for student enquiries. *Journal of Physics: Conference Series*, *1228*(1). https://doi.org/10.1088/1742-6596/1228/1/012060

Tavakoli, P. (2009). Investigating task difficulty: Learners' and teachers' perceptions. *International Journal of Applied Linguistics*, *19*(1), 1–25. https://doi.org/10.1111/j.1473-4192.2009.00216.x

Toxtli, C., Monroy-Hernández, A., & Cranshaw, J. (2018). Understanding chatbot-mediated task management. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3173574.3173632

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, *151*, 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

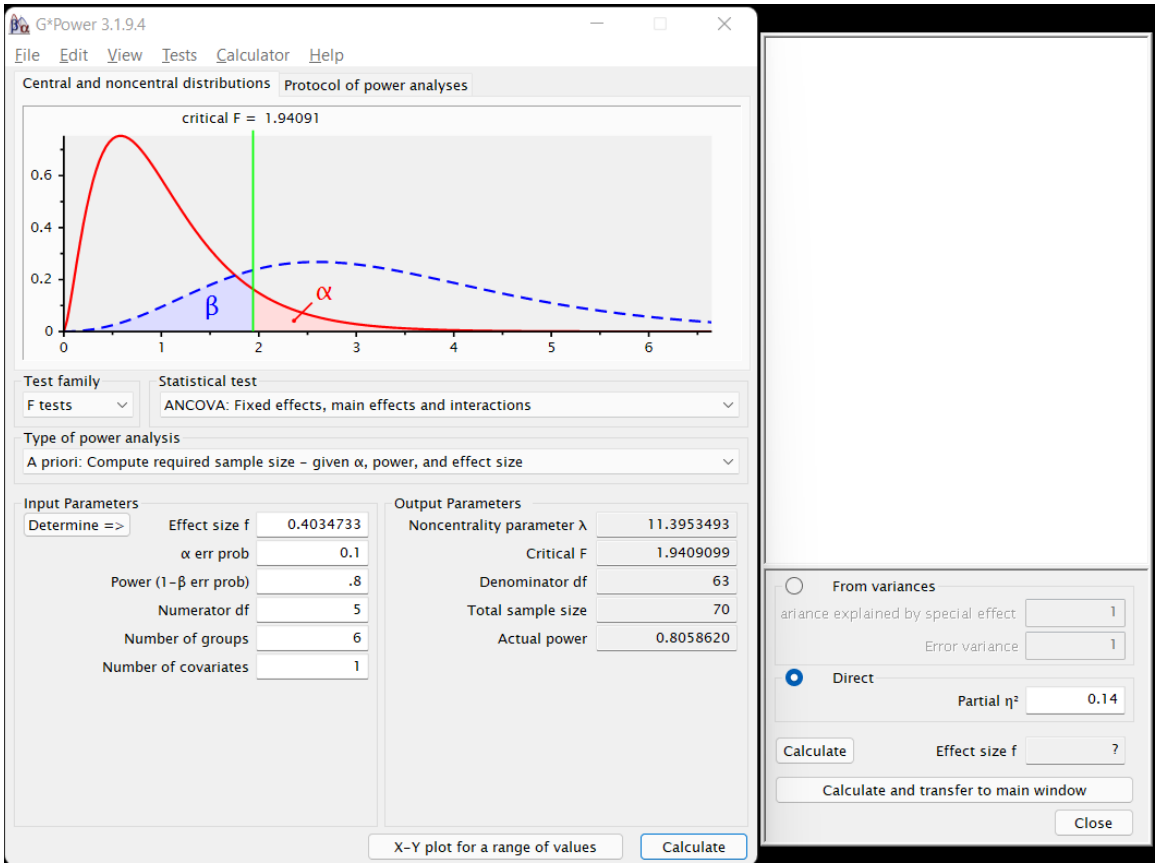
Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). Unified theory of acceptance and use of Technology. *PsycTESTS Dataset*. https://doi.org/10.1037/t57185-000

Weingart, P. (2019). A chatbot dialogue model: Understanding human-chatbot collaboration in a complex task environment. *14th International Conference on Wirtschaftsinformatik in Siegen*. https://www.alexandria.unisg.ch/256977/1/student%20paper%20weingart.pdf

APPENDIX A

G*POWER ANALYSIS


G*Power 3.1.9.4
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 1.94091
0.6
0.4
0.2
0
0 1 2 3 4 5 6
β
α
Test family
F tests
Statistical test
ANCOVA: Fixed effects, main effects and interactions
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Determine =>
Effect size f 0.4034733
α err prob 0.1
Power (1–β err prob) .8
Numerator df 5
Number of groups 6
Number of covariates 1
Output Parameters
Noncentrality parameter λ 11.3953493
Critical F 1.9409099
Denominator df 63
Total sample size 70
Actual power 0.8058620
X-Y plot for a range of values
Calculate
From variances
ariance explained by special effect 1
Error variance 1
Direct
Partial η² 0.14
Calculate
Effect size f ?
Calculate and transfer to main window
Close

APPENDIX B

DEMOGRAPHIC SECTION

D1 What is your age?

- 18 - 24 (1)
- 25 - 34 (2)
- 35 - 44 (3)
- 45 - 54 (4)
- 55 - 64 (5)
- 65 - 74 (6)
- 75 - 84 (7)
- 85 or older (8)

D2 What is your level of education?

- Less than high school (1)
- High school graduate (2)
- Some college (3)
- 2 year degree (4)
- 4 year degree (5)
- Professional degree (6)
- Doctorate (7)

D3 How often do you use a computer?

- Never (1)
- Once a week (2)
- 2-3 times a week (3)
- 4-6 times a week (4)
- Daily (5)

D4 How confident do you feel in your ability to use a computer?

- Not confident at all (1)
- Somewhat confident (2)
- Confident (3)
- Very confident (4)

D5 How would you best describe your experience with chatbots?

- I have never interacted with a chatbot (or I do not know if I have) (1)
- I rarely interact with chatbots (2)
- I sometimes interact with chatbots (3)
- I interact with chatbots often (4)

D6 How would you best describe your feelings about chatbots?

- I hate chatbots (1)
- I dislike chatbots (2)
- I have neutral feelings toward chatbots (3)
- I like chatbots (4)
- I love chatbots (5)

APPENDIX C

USER EXPERIENCE INSTRUMENT

Table 2

User Experience Instrument

Statement	Metric Measured
1) Using this system improves the quality of the work I do.	Perceived Usefulness
2) Using this system gives me greater control over my work.	Perceived Usefulness
3) This system enables me to accomplish tasks more quickly.	Perceived Usefulness
4) This system supports critical aspects.	Perceived Usefulness
5) This system increases my productivity.	Perceived Usefulness
6) This system improves my task performance.	Perceived Usefulness
7) This system allows me to accomplish more work than would otherwise be possible.	Perceived Usefulness
8) This system enhances my effectiveness on the task.	Perceived Usefulness
9) This system makes it easier to do my task.	Perceived Usefulness
10) Overall, I find this system useful in my task.	Perceived Usefulness
11) The system's capabilities meet my requirements.	Usability
12) Using this system is a frustrating experience.	Usability
13) This system is easy to use.	Usability
14) I have to spend too much time correcting things with this system.	Usability
15) Overall, I am satisfied with the system.	Satisfaction
16) I would use this system again.	Satisfaction

Note. All statements were followed by a Likert-scale with the following possible participant responses: (1) Strongly disagree, (2) Disagree, (3) Somewhat disagree, (4) Somewhat agree, (5) Agree, (6) Strongly agree

APPENDIX D

RULE-BASED AGENT SCRIPTS

Repeat

Simple

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?
- OK, what is the pin number associated with your Bluewidth account?

  I'm sorry, I couldn't find your account. What is your pin number?
- Just a moment…
- I found your account. You are not renting any wifi equipment from us currently. Do you already own personal wifi equipment? Requesting new equipment will replace this equipment.
- What wifi equipment do you already own?
- `I don't quite understand. What wifi equipment do you already own?`
- I see you have an internet router and an internet modem. Would you like to rent this same equipment from us?
- What is the specific name of your wifi equipment?
- `I don't quite understand. What is the specific name of your wifi equipment?`
- Just a moment while I search our inventory for this equipment…
- I found your equipment. We have this equipment available. I will have a representative hold these items for you to pick up at the nearest Bluewidth Internet location. You can pick them up during regular business hours. The location nearest to your listed address is:
  1345 S. Bridge Rd,
  Tempe, AZ, 85281
- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request right now. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.
- Have a great day! Thanks for choosing Bluewidth Internet!

`I don't quite understand.`

I'm sorry, I don't quite understand

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I'm sorry I can't help you with that request.

Yes

No

I'm sorry I cannot process your request.

I'm sorry I cannot process your request. Can you rephrase your response?

Our regular business hours are 9am-6pm on Monday-Friday

I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

Complex

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?

- OK, what is the pin number associated with your Bluewidth account?

  I'm sorry, I couldn't find your account. What is your pin number?

- Just a moment…

- I found your account. Just a moment while I look over your statements and most recent charge history…

- OK, I see here your last transaction was on November 23rd 2021, where a payment was made for a new internet modem. The modem is a Netgear 32 x 8 Docsis 3.1 modem. Does this sound correct?

  This is the only transaction on record. Are you sure this is not the correct transaction?

- OK, is this the transaction in which you are requesting a refund?

  This is your only transaction on record. Are you sure this is not the correct transaction?

- What is your reason for requesting a refund?

- I don't quite understand. What is your reason for requesting a refund?

- OK, your response was recorded. I will need more information so I can file a claim for your refund. Can you confirm your address?

- I don't quite understand. Can you confirm your address?

- OK, are you currently using and/or renting the item that you are requesting a refund for?

- What is the model number of the item you need a refund for (if you have it)?

- I don't quite understand, what is the model # of the item you need a refund for (if you have it)?

- OK, will you be needing a replacement for the item that you are requesting a refund for?

- OK, would you like your refund to be store credit or US dollars to the debit/credit card that was used to pay your charge?

    I don't quite understand, would you like your refund to be store credit or US dollars to the debit/credit card that was used to pay your charge?

    OK, I will refund your account the store credit for this transaction.

    Is there anything else I can help you with today?

    Have a great day! Thanks for choosing Bluewidth Internet!

- OK, can you please confirm the last 4 digits of the debit/credit that was used to pay your charge?

- I don't quite understand, can you please confirm the last 4 digits of the debit/credit that was used to pay your charge?

- Thank you. I have created a new claim for your refund. A representative will review your claim and contact you to go over your request. They will give you further steps to follow in regard to your refund and item return (if applicable).

- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request right now. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

- Have a great day! Thanks for choosing Bluewidth Internet!

`I don't quite understand`

I'm sorry, I don't quite understand

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I'm sorry I can't help you with that request.

Yes

No

I'm sorry I cannot process your request.

I'm sorry I cannot process your request. Can you rephrase your response?

Our regular business hours are 9am-6pm on Monday-Friday

I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

— —

Confirmation

Simple

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?
- OK, what is the pin number associated with your Bluewidth account?

  I'm sorry, I couldn't find your account. What is your pin number?

- Just a moment…
- I found your account. You are not renting any wifi equipment from us currently. Do you already own personal wifi equipment? Requesting new equipment will replace this equipment.

- What wifi equipment do you already own?
- What I think you said is, you own a router, but not a modem. Is this correct?
- What I think you said is, you own a router and a modem. Is this correct?
- Would you like to rent this same equipment from us?
- OK, what is the specific name of your wifi equipment?
- What I think you said is, you own a Netgear CM700 Docsis 3.0 Cable Modem and a Netgear Nighthawk AX6 WLAN Router RAX50. Is that right?
- What I think you said is, you own a Netgear 32 x 8 Docsis 3.1 and a Netgear Nighthawk AX6 WLAN Router RAX50. Is that right?
- Just a moment while I search our inventory for this equipment…
- I found your equipment. We have this equipment available. I will have a representative hold these items for you to pick up at the nearest Bluewidth Internet location. You can pick them up during regular business hours. The location nearest to your listed address is:
  1345 S. Bridge Rd,
  Tempe, AZ, 85281
- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request right now. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day!
- Have a great day! Thanks for choosing Bluewidth Internet!

What I think you said is,

I'm sorry, I don't quite understand

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I’m sorry I can’t help you with that request.

Yes

No

I'm sorry I cannot process your request.

I'm sorry I cannot process your request. Can you rephrase your response?

Our regular business hours are 9am-6pm on Monday-Friday

I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

Complex

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?
- OK, what is the pin number associated with your Bluewidth account?

  I’m sorry, I couldn’t find your account. What is your pin number?

- Just a moment…
- I found your account. Just a moment while I look over your statements and most recent charge history...

- OK, I see here your last transaction was on November 23rd 2021, where a payment was made for a new internet modem. The modem is a Netgear 32 x 8 Docsis 3.1 modem. Does this sound correct?

  This is the only transaction on record. Are you sure this is not the correct transaction?

- OK, is this the transaction in which you are requesting a refund?

  This is the only transaction on record. Are you sure this is not the correct transaction?

- What is your reason for requesting a refund?

- `What I think you said is, your modem was never delivered. Is this correct?`

- What I think you said is, you were mistakenly charged for this item. Is this correct?

- OK, your response was recorded. I will need more information so I can file a claim for your refund. Can you confirm your address?

- What I think you said is, your address is 7654 W. Kennard St., Flint, MI, 48506. Is that right?

- What I think you said is, your address is 7654 W. Kennard St. Tempe, AZ, 85281. Is this correct?

- OK, are you currently using and/or renting the item that you are requesting a refund for?

- What is the model number of the item you need a refund for (if you have it)?

- What I think you said is, the model number is 120673883. Is this correct?

- What I think you said is, the model number is 120673983. Is this correct?

- OK, will you be needing a replacement for the item that you are requesting a refund for?

- OK, would you like your refund to be store credit or US dollars to the debit/credit card that was used to pay your charge?

  What I think you said is, refund by debit/credit card. Is that right?

  OK, I will refund your account the store credit for this transaction.

  Is there anything else I can help you with today?

  Have a great day! Thanks for choosing Bluewidth Internet!

- OK, can you please confirm the last 4 digits of the debit/credit that was used to pay your charge?

- What I think you said is, 4393. Is that right?

- What I think you said is, 4294. Is that right?

- Thank you. I have created a new claim for your refund. A representative will review your claim and contact you to go over your request. They will give you further steps to follow in regard to your refund and item return (if applicable).

- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request right now. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

- Have a great day! Thanks for choosing Bluewidth Internet!

What I think you said is,

I'm sorry, I don't quite understand

I am sorry I could not find your account. Please try to send your pin again.

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I'm sorry I can't help you with that request.

Yes

No

I'm sorry I cannot process your request.

I'm sorry I cannot process your request. Can you rephrase your response?

Our regular business hours are 9am-6pm on Monday-Friday

I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

— —

Options

Simple

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?
- OK, what is the pin number associated with your Bluewidth account?

  I'm sorry, I couldn't find your account. What is your pin number?
- I found your account. You are not renting any wifi equipment from us currently. Do you already own personal wifi equipment? Requesting new equipment will replace this equipment.
- What wifi equipment do you already own?

- I am not quite sure if I understood. Which of these do you already own:
  1) An internet router,
  2) An internet modem,
  3) Both an internet router and an internet modem,
  4) None of the above

- I see you have an internet router and an internet modem. Would like to rent this same equipment from us?

- What is the specific name of your wifi equipment?

- I am not quite sure if I got that correct. Which modem did you want:
  1) Netgear 32 x 8 Docsis 3.1,
  2) Netgear CM700 Docsis 3.0 Cable Modem
  3) Netgear Nighthawk Docsis 3.1 Cable Modem
  4) None of these

- Just a moment while I search our inventory for this equipment…

- I found your equipment. We have this equipment available. I will have a representative hold these items for you to pick up at the nearest Bluewidth Internet location. You can pick them up during regular business hours. The location nearest to your listed address is:
  1345 S. Bridge Rd,
  Tempe, AZ, 85281

- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request right now. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

- Have a great day! Thanks for choosing Bluewidth Internet!

I am not quite sure if I understood.

I am not quite if I got that correct.

I'm sorry, I don't quite understand

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I'm sorry I can't help you with that request.

Yes

No

I'm sorry I cannot process your request.

I'm sorry I cannot process your request. Can you rephrase your response?

Our regular business hours are 9am-6pm on Monday-Friday

I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

Complex

- Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?

- OK, what is the pin number associated with your Bluewidth account?

  I'm sorry, I couldn't find your account. What is your pin number?

- Just a moment…
- I found your account. Just a moment while I look over your statements and most recent charge history...

- OK, I see here your last transaction was on November 23rd 2021, where a payment was made for a new internet modem. The modem is a Netgear 32 x 8 Docsis 3.1 modem. Does this sound correct?

  This is the only transaction on record. Are you sure this is not the correct transaction?

- OK, is this the transaction in which you are requesting a refund?

  This is the only transaction on record. Are you sure this is not the correct transaction?

- What is your reason for requesting a refund?

- ```
  I am not quite sure if I understood. Which of these is your
  reason for requesting a refund:
  1) Item was not as described,
  2) Item was damaged/malfunctioning
  3) Item was not the correct item
  4) Item was not ordered
  5) None of the above
  ```
  ```

- OK, your response was recorded. I will need more information so I can file a claim for your refund. Can you confirm your address?

- I am not quite sure if I got that correct. Which of these is your address:
  1) 7654 W. Kennard St., Tempe, AZ, 85281
  2) 7654 W. Kennard St., Flint, MI, 48506
  3) 7654 W. Kennard Rd, Mesa, AZ 85204
  4) None of the above

- OK, are you currently using and/or renting the item that you are requesting a refund for?

- What is the model number of the item you need a refund for (if you have it)?

- I am not quite sure if I got that correct. Which of these is the model number:
  1) 120673984
  2) 121673983
  3) 120673983
  4) None of the above

- OK, will you be needing a replacement for the item that you are requesting a refund for?

- OK, would you like your refund to be store credit or US dollars to the debit/credit card that was used to pay your charge?

  I am not quite sure if I got that correct. How do you want your refund:
  1) Store Credit
  2) US Dollars on your debit/credit card
  3) None of the above

  OK, I will refund your account the store credit for this transaction.

  Is there anything else I can help you with today?

  Have a great day! Thanks for choosing Bluewidth Internet!

- OK, can you please confirm the last 4 digits of the debit/credit card that was used to pay your charge?

- I am not quite sure if I got that correct. Which of these is the last four digits of the debit/credit card that was used to pay your charge:
  1) 4293
  2) 4294
  3) 4394
  4) None of the above

- Thank you. I have created a new claim for your refund. A representative will review your claim and contact you to go over your request. They will give you further steps to follow in

regard to your refund and item return (if applicable).

- Is there anything else I can help you with today?

  I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

- Have a great day! Thanks for choosing Bluewidth Internet!

I am not quite sure if I understood.

I am not quite sure if I got that correct.

I'm sorry, I don't quite understand

Are you still there?

Please standby while I process your request

Just a moment…

Let me try to answer your question. Just a moment…

I'm sorry I can't find the answer to your question

I'm sorry I can't help you with that request.

Yes

No

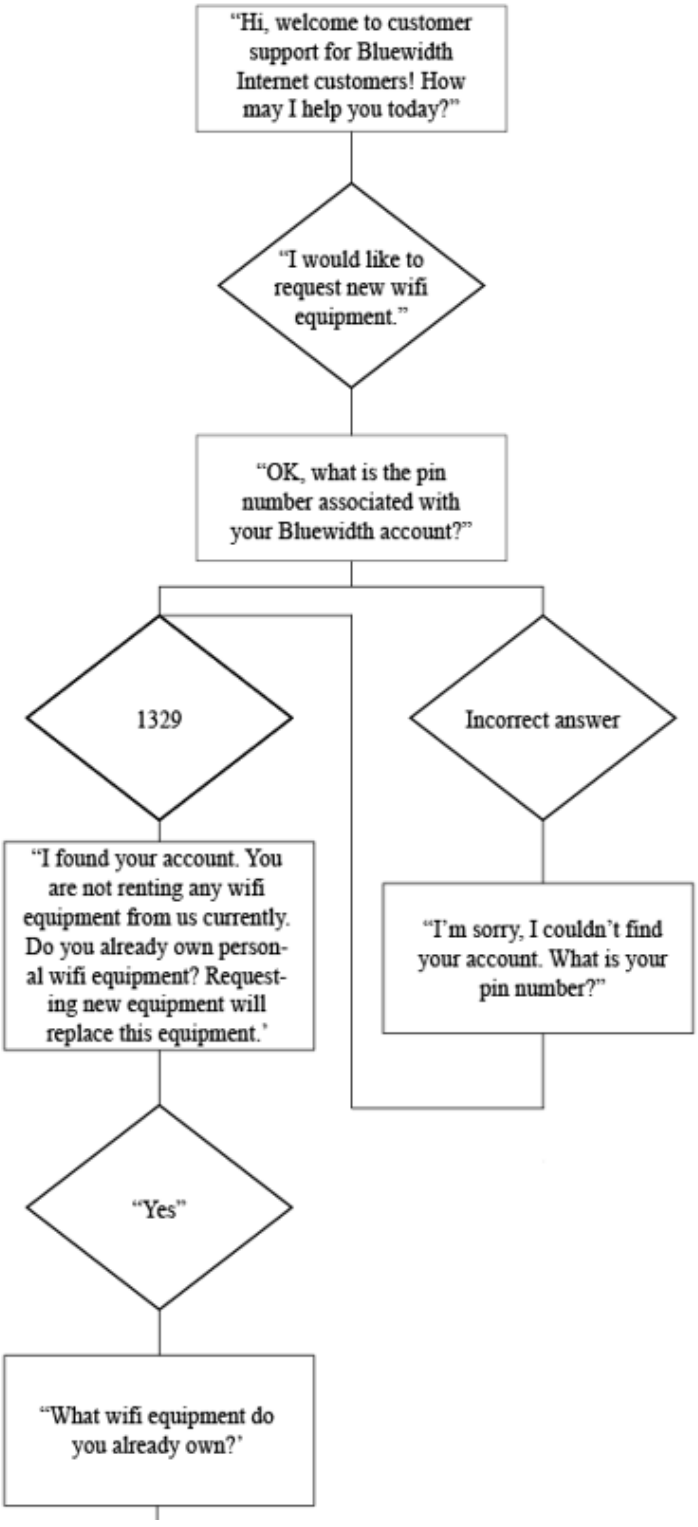
I'm sorry I cannot process your request.

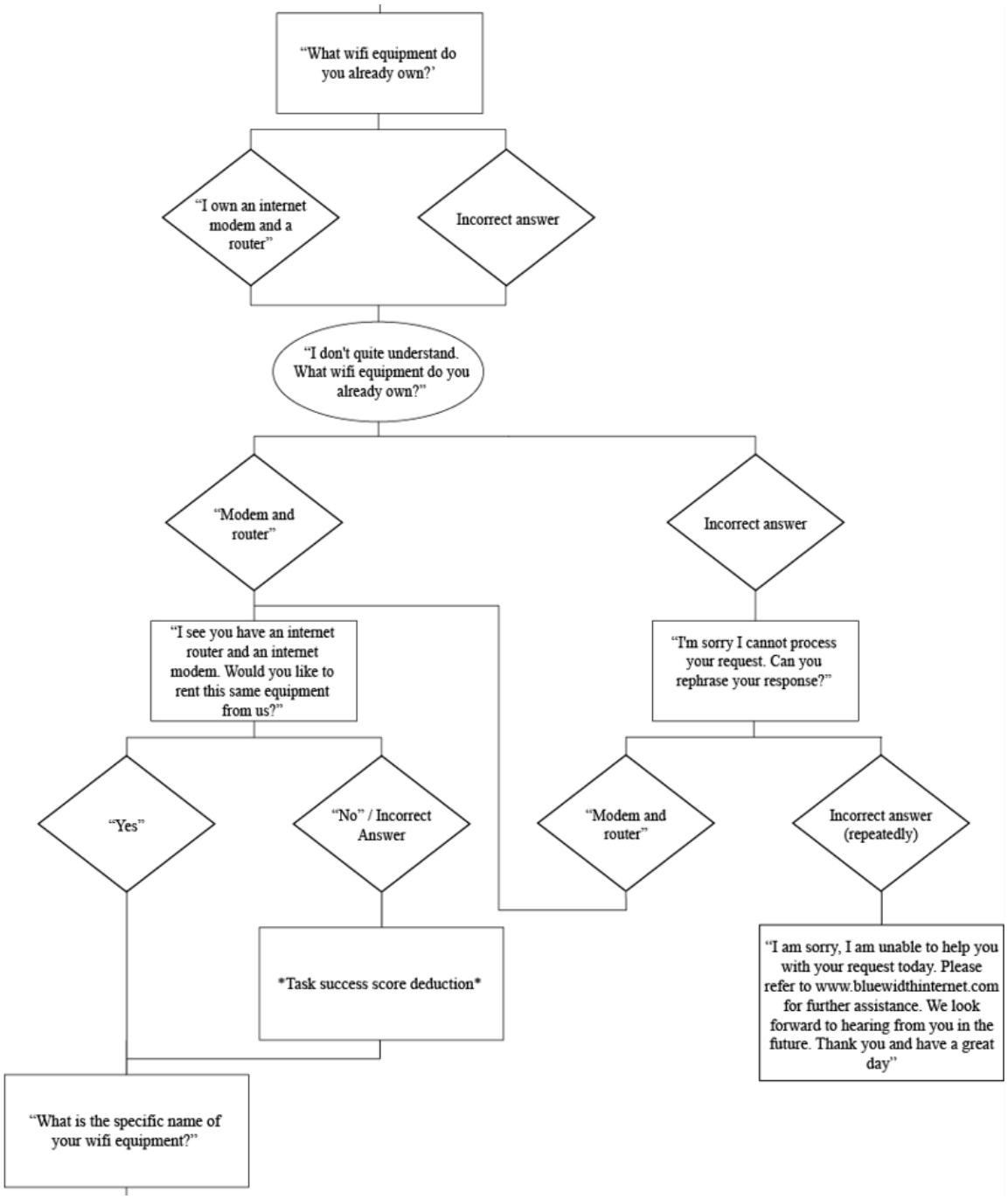
I'm sorry I cannot process your request. Can you rephrase your response?
Our regular business hours are 9am-6pm on Monday-Friday

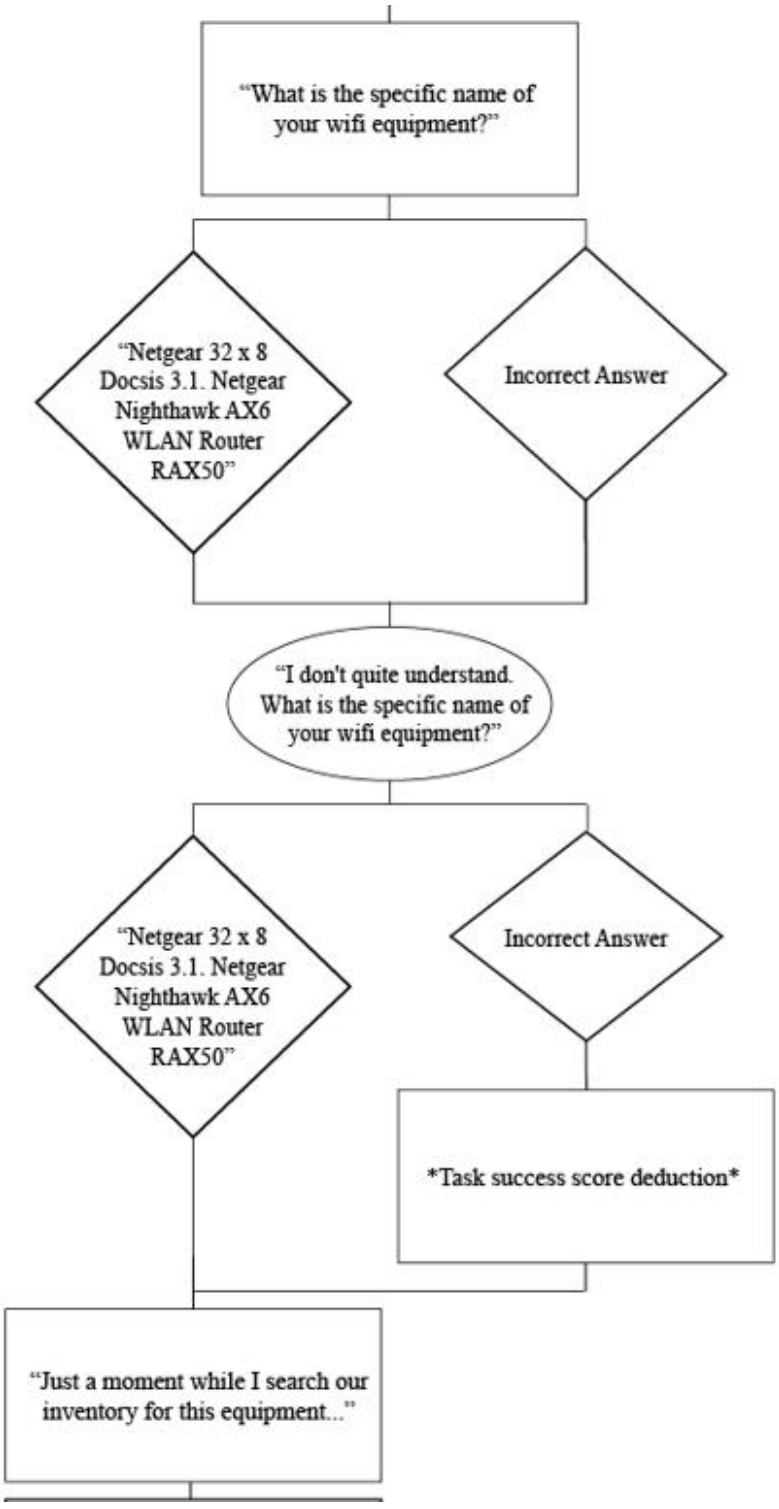
I am sorry. I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day.

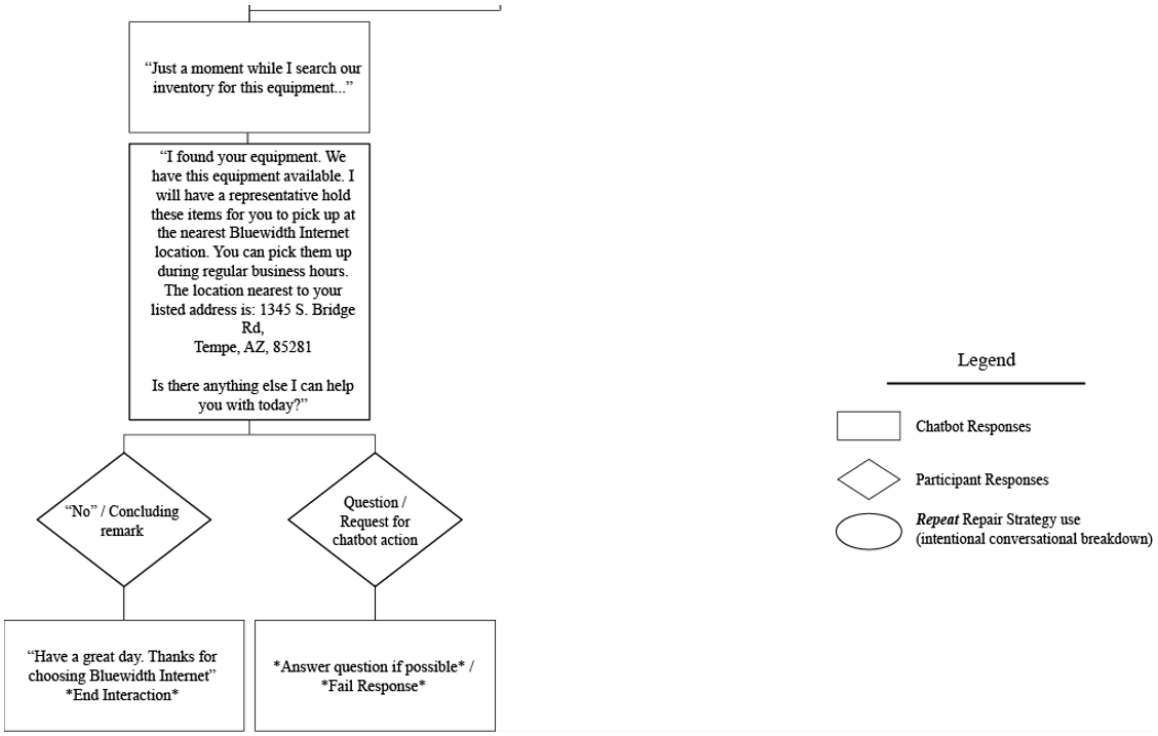
APPENDIX E

HUMAN-CHATBOT CONVERSATIONAL FLOW DIAGRAM


"Hi, welcome to customer support for Bluewidth Internet customers! How may I help you today?"
"I would like to request new wifi equipment."
"OK, what is the pin number associated with your Bluewidth account?"
1329
Incorrect answer
"I found your account. You are not renting any wifi equipment from us currently. Do you already own personal wifi equipment? Requesting new equipment will replace this equipment.'
"I'm sorry, I couldn't find your account. What is your pin number?"
"Yes"
"What wifi equipment do you already own?'


"What wifi equipment do you already own?'
"I own an internet modem and a router"
Incorrect answer
"I don't quite understand. What wifi equipment do you already own?"
"Modem and router"
Incorrect answer
"I see you have an internet router and an internet modem. Would you like to rent this same equipment from us?"
"I'm sorry I cannot process your request. Can you rephrase your response?"
"Yes"
"No" / Incorrect Answer
"Modem and router"
Incorrect answer (repeatedly)
Task success score deduction
"I am sorry, I am unable to help you with your request today. Please refer to www.bluewidthinternet.com for further assistance. We look forward to hearing from you in the future. Thank you and have a great day"
"What is the specific name of your wifi equipment?"


"What is the specific name of your wifi equipment?"
"Netgear 32 x 8 Docsis 3.1. Netgear Nighthawk AX6 WLAN Router RAX50"
Incorrect Answer
"I don't quite understand. What is the specific name of your wifi equipment?"
"Netgear 32 x 8 Docsis 3.1. Netgear Nighthawk AX6 WLAN Router RAX50"
Incorrect Answer
Task success score deduction
"Just a moment while I search our inventory for this equipment..."


"Just a moment while I search our inventory for this equipment..."
"I found your equipment. We have this equipment available. I will have a representative hold these items for you to pick up at the nearest Bluewidth Internet location. You can pick them up during regular business hours. The location nearest to your listed address is: 1345 S. Bridge Rd, Tempe, AZ, 85281
Is there anything else I can help you with today?"
"No" / Concluding remark
Question / Request for chatbot action
"Have a great day. Thanks for choosing Bluewidth Internet" *End Interaction*
Answer question if possible / *Fail Response*
Legend
Chatbot Responses
Participant Responses
Repeat Repair Strategy use (intentional conversational breakdown)

APPENDIX F

SIMPLE TASK INFORMATION DOCUMENT

Study Context

You will be helping Bluewidth Internet, an up and coming internet service provider, by speaking to their automated customer service chatbot. You will be given a task to complete by speaking to their chatbot representative while playing the role of a customer of Bluewidth Internet. Your effort in this study will prove valuable to improving their customer support chatbot. You will also be scored on your ability to complete this task and debriefed on your performance at the end. **Do you have any questions?**

Keep in mind that the Bluewidth chatbot is still in development and may not be able to answer promptly or effectively in all cases. Any questions to the chatbot may be met with limited responses or no response at all. The experimenter will not be available to answer questions once your task begins but may intervene on behalf of the chatbot if necessary. Thank you for your participation.

Participant Instructions:

- You will be communicating with the Bluewidth Internet Chatbot over the Zoom chat to complete a task that the experimenter will give to you in just a moment
- You will be given 2 minutes to read over the information about your task
- You will then interact with the chatbot to complete the task
- You will be scored on your ability to complete the task successfully
- You will answer questions about your experience after the conclusion of the task

Task Information

Setting:

You are currently a customer of Bluewidth Internet, an internet service provider, that you pay on a monthly basis for your internet access.

Your Task:

Talk to Bluewidth Internet customer services to **request new wifi equipment**.

Begin the conversation by saying:

"I would like to request new wifi equipment."

Additional information for your understanding:

- You need a new internet modem and router
- The wifi equipment that you own is broken
- You want to rent the exact same wifi equipment that you already own from Bluewidth as replacement for your broken equipment
- The address of your internet services is 7654 W. Kennard St., Tempe, AZ, 85281
- The broken wifi equipment that you own is an:
    - Internet Modem
        - Modem Name: Netgear 32 x 8 Docsis 3.1
        - Model #: 120673983
    - Internet Router
        - Router Name: Netgear Nighthawk AX6 WLAN Router RAX50
        - Model #: 268900563
- Your PIN number for your account is: 1329

APPENDIX G

COMPLEX TASK INFORMATION DOCUMENT

Study Context

You will be helping Bluewidth Internet, an up and coming internet service provider, by speaking to their automated customer service chatbot. You will be given a task to complete by speaking to their chatbot representative while playing the role of a customer of Bluewidth Internet. Your effort in this study will prove valuable to improving their customer support chatbot. You will also be scored on your ability to complete this task and debriefed on your performance at the end. **Do you have any questions?**

Please keep in mind that the Bluewidth chatbot is still in development and may not be able to answer promptly or effectively in all cases. Any questions to the chatbot may be met with limited responses or no response at all. The experimenter will not be available to answer questions once your task begins but may intervene on behalf of the chatbot if necessary. Thank you for your participation.

Participant Instructions:

- You will be communicating with the Bluewidth Internet Chatbot over the Zoom chat to complete a task that the experimenter will give to you in just a moment
- You will be given 2 minutes to read over the information about your task
- You will then interact with the chatbot to complete the task
- You will be scored on your ability to complete the task successfully
- You will answer questions about your experience after the conclusion of the task

Task Information

Setting:

You are currently a customer of Bluewidth Internet, an internet service provider, that you pay on a monthly basis for your internet access.

Your Task:

Talk to Bluewidth Internet customer services to **get a refund for a charge that was applied to your account that should not have been there.**

Begin the conversation by saying:

"I need a refund because my account was charged and I am not sure why."

Additional information for your understanding:

- You have the Speedy 500 Internet Plan, where you pay $79.99 for 500 mbps download speed
- You pay an additional $14.99 a month for your wifi equipment
- You recently had to trade in your previous Bluewidth modem for another one
    - Your previous modem was malfunctioning and it was not because of you or your use of it so it did not cost you to do this
    - You returned the malfunctioning modem on 11/23/21 (details below):
        - Modem Name: Netgear CM700 Docsis 3.0 Cable Modem
        - Monthly Price: $10.99
    - Your replacement modem was a Netgear 32 x 8 Docsis 3.1
- Bluewidth seemed to have charged you $79.99 and you paid it with your credit card not knowing it was not supposed to be applied to your account
    - You want this $79.99 charge refunded to your credit card
    - The credit card you used and want your refund on is:
        - Visa - 4590 3902 3893 4294

- The address associated with your internet services:
    - 7654 W. Kennard St., Tempe, AZ, 85281
- The equipment that you currently rent from Bluewidth is:
    - Internet Modem
        - Modem Name: Netgear 32 x 8 Docsis 3.1
        - Model #: 120673983
        - Monthly Price: $10.99
    - Internet Router
        - Router Name: Netgear Nighthawk AX6 WLAN Router RAX50
        - Model #: 268900563
        - Monthly Price: $11.99
- You have a promotion package that only charges $14.99 monthly for both the modem and router that you rent from Bluewidth.
- Your PIN number for your account is: 1329