Highlights

The Way You Assess Matters: User Interaction Design of Survey Chatbots for Mental Health

Yucheng Jin, Li Chen, Xianglin Zhao, Wanling Cai

- This work investigates how the interaction design of psychological assessment with closed-ended questions could influence user responses to open-ended questions in a survey chatbot for mental health.
- An empirical study shows the significant effects of *interaction style* (form-based vs. conversation-based) on user-perceived assessment credibility and self-awareness.
- A structural equation model illustrates the mediating role of perceived assessment credibility in the effects of psychological assessment design on user responses to the subsequent open-ended questions.

The Way You Assess Matters: User Interaction Design of Survey Chatbots for Mental Health

Yucheng Jin[a], Li Chen[a], Xianglin Zhao[a], Wanling Cai[a]

[a]*Department of Computer Science, Hong Kong Baptist University, Hong Kong, China*

Abstract

The global pandemic has pushed human society into a mental health crisis, prompting the development of various chatbots to supplement the limited mental health workforce. Several organizations have employed mental health survey chatbots for public mental status assessments. These survey chatbots typically ask closed-ended questions (Closed-EQs) to assess specific psychological issues like anxiety, depression, and loneliness, followed by open-ended questions (Open-EQs) for deeper insights. While Open-EQs are naturally presented conversationally in a survey chatbot, Closed-EQs can be delivered as embedded forms or within conversations, with the length of the questionnaire varying according to the psychological assessment. This study investigates how the *interaction style* of Closed-EQs and the *questionnaire length* affect user perceptions regarding survey credibility, enjoyment, and self-awareness, as well as their responses to Open-EQs in terms of quality and self-disclosure in a survey chatbot. We conducted a 2 (*interaction style*: form-based vs. conversation-based) × 3 (*questionnaire length*: short vs. middle vs. long) between-subjects study (N=213) with a loneliness survey chatbot. The results indicate that the form-based interaction significantly enhances the perceived credibility of the assessment, thereby improving response quality and self-disclosure in subsequent Open-EQs and fostering self-awareness. We discuss our findings for the interaction design of psychological assessment in a survey chatbot for mental health.

Keywords: Chatbots, survey design, open-ended questions, psychological assessment, self-disclosure, mental health, loneliness

1. Introduction

The rise of mental health issues among young adults has become a significant public health challenge [1, 2, 3, 4], further intensified by the global pandemic's

Email address: `yuchengjin@hkbu.edu.hk` (Yucheng Jin)

impact on various aspects of life [5, 6, 7]. Early detection and intervention are crucial for providing targeted support and treatments [8, 9]. With the rapid advancement in artificial intelligence (AI), several organizations, including universities, hospitals, and public sectors, have begun utilizing mental health survey chatbots for conducting psychological assessments to determine individuals' mental states and needs [10, 11, 12]. Compared with traditional web-based surveys, chatbot surveys have demonstrated advantages in response rate, user engagement, and response quality due to the natural conversation and interactive features [13, 14].

Mental health surveys typically contain two primary types of questions [15, 16]: *closed-ended questions* (Closed-EQs) often based on psychological scales like the UCLA Loneliness Scale consisting of twenty Closed-EQs [17], and open-ended questions (Open-EQs) that delve into deeper individual insights [16], promoting spontaneous and less biased responses [15]. Research in web surveys has revealed correlations between responses to Closed-EQs and subsequent Open-EQs. For example, participants dissatisfied with job or e-services through Closed-EQs tended to disclose more details about negative feelings in subsequent Open-EQs [18, 19, 20]. However, little work has investigated *if* and *how* the design choices of Closed-EQs influence user responses to Open-EQs, particularly in mental health survey chatbots. Existing work has primarily investigated leveraging a chatbot to respectively improve the response quality of Closed-EQs or Open-EQs [14, 21]. Our research aims to bridge the gap by exploring the effects of two prominent design factors (i.e., *interaction style* and *questionnaire length*) of a psychological assessment with Closed-EQs on user responses to the follow-up Open-EQs in a mental health survey chatbot.

The *interaction style* and *questionnaire length* are two crucial design factors of Closed-EQs [22, 23]. Prior studies have shown that, compared to conventional form-based interactions on webpages, employing conversation-based interactions can enhance the quality of responses to Closed-EQs [14]. Additionally, research has demonstrated that the *questionnaire length* can influence participation and completion rate [23, 24], as well as the response quality [25]. In our study, we experimented with both form-based and conversation-based interactions in our chatbot's psychological assessment. The manipulation of questionnaire length is based on the three validated versions of the UCLA loneliness scale [17], including short (three items), middle (ten items), and long (twenty items), respectively. This led to a 2 (interaction style: form-based vs. conversation-based) $\times$ 3 (questionnaire length: short vs. middle vs. long) between-subjects study, enabling us to address the following four research questions with empirical evidence.

RQ1: How does the *interaction style* of an assessment influence the users' perceptions of a mental health survey chatbot (i.e., w.r.t. enjoyment, assessment credibility, and self-awareness)?

RQ2: How does the *interaction style* of an assessment influence user responses to the follow-up Open-EQs (i.e., w.r.t. response quality and self-disclosure) in a mental health survey chatbot?

RQ3: How does the *questionnaire length* of an assessment influence the users' perceptions of a mental health survey chatbot?

RQ4: How does the *questionnaire length* of an assessment influence user responses to the follow-up Open-EQs in a mental health survey chatbot?

Our study provides practical design implications to designers of survey chatbots for mental health. To the best of our knowledge, this is the first study that empirically analyzes how psychological assessment design influences user responses to Open-EQs within a mental health survey chatbot. Consequently, the contributions of our work are three-fold:

1. **Empirical evidence of the effects of psychological assessment (with Closed-EQs) design on user responses to the follow-up Open-EQs in a survey chatbot for mental health.** Our findings reveal the effective design choices for the psychological assessment that could motivate respondents to provide quality responses and stimulate deep self-disclosure in Open-EQ.

2. **Analysis of the causal relationship between the design factors of psychological assessment and the measures for user responses to Open-EQs.** We employed a structural equation model (SEM) to identify how users' *perceived assessment credibility*, as a mediator, links psychological assessment design factors to the critical metrics of user responses to Open-EQs such as response quality and self-disclosure.

3. **Design recommendations of psychological assessment in a survey chatbot for mental health.** Based on our findings, we present several practical design recommendations. For instance, form-based interaction is preferable for psychological assessments, as it leads to a higher perceived assessment credibility compared to conversation-based interaction.

2. Related work

2.1. Loneliness Among Young Adults and Its Measurement

Loneliness is a common distressing feeling that is closely associated with adverse mental health states, such as depression and anxiety [26, 27, 28, 29]. Young people are more susceptible to loneliness compared to other age groups, due to a dramatic increase in socioemotional demands at their unique life stage [30, 31]. The social restrictions imposed to control the spread of COVID-19 have notably diminished social contact for the youth, exacerbating their feelings of loneliness and leading to increased psychological distress [32, 33, 34]. For example, following the outbreak of COVID-19, up to 60% of young adults in America have reported symptoms indicative of psychological distress [35].

Early detection and intervention of loneliness are crucial for young adults, as these steps can help them mitigate its long-term effects on their mental health and support them in establishing healthier social connections and networks [34]. When measuring loneliness, the UCLA Loneliness Scale and its related shorter forms are widely acknowledged and recommended as the primary tools for assessing loneliness [36]. As for intervention strategies, recent studies highlight the effectiveness of chatbots as an innovative method to offer essential social support. They serve as a valuable tool in fostering users' reflection on their emotional self-awareness, social awareness, and interpersonal relationships, which will be described in detail in the following section. Considering the context of our study and the prevalence of loneliness among young adults, particularly in the era of the COVID-19 pandemic, our study has focused on loneliness in our psychometric assessments.

2.2. Chatbots for Mental Health

Chatbots have great potential to promote mental health by conversing with users to provide psychological assessment, training, and therapy [10]. For example, Woebot [1] and Wysa [2] are representative chatbots for mental health; and their efficacy has been proven by clinical research [37, 38]. To assess users' emotional state or the severity of a specific mental health issue, some chatbots ask questions based on some well-known psychological scales, such as PHQ-9 Depression Test Questionnaire [39] and Generalized Anxiety Disorder Assessment (GAD-7) [40]. Performing assessment in a chatbot tends to be an effective way to collect mental health data, comparable to physical interviews in terms of response rate [11]. Based on users' responses to the assessment questions, chatbots provide empathetic responses, emotion diary, mindfulness exercises, and goal setting to help users cope with mental health issues [41, 42]. Existing Human-Computer

[1]https://woebothealth.com/
[2]https://www.wysa.io/

Interaction (HCI) research in mental health chatbots focuses on improving conversation skills to demonstrate compassion and empathy [43, 44] and promote user self-disclosure [45, 46], and integrating various practices for mental health (e.g., expressive writing [47], motivational interview [48], and social support [49]) into chatbots. However, little work has studied the psychological assessment design and its impacts on user responses in a survey chatbot for mental health.

2.3. Design for Online Psychological Assessment

The computer-based psychological assessment allows users to employ valid psychological scales to quickly gauge a specific mental health aspect such as loneliness, anxiety, and depression [50]. The psychological assessment is often performed by asking users to answer a set of closed-ended questions, similar to the questionnaire. Interaction style and questionnaire length are two major design factors that could influence the participation rate and response quality of a questionnaire [51, 52, 53, 54, 55]. Therefore, we mainly review the related work of *interaction style* and *questionnaire length* that we have manipulated in our study.

2.3.1. Interaction Style

Prior work shows mixed effects of the interaction style on user responses to questionnaires. The ways of showing the questions (multiple short pages vs. a long scrollable page) and adding more interactive elements (i.e., pop-up menus, button scales, and numerical labeling) do not yield a significant difference in user response behavior [56, 57]. In contrast, compared to the item-by-item questions, showing questions in a matrix may increase non-response items [58]. Additionally, interaction style could affect users' perceived credibility of information on the web [59]. Within a chatbot, some social characteristics (e.g., proactivity and conscientiousness) could also influence users' perceived credibility [60]. As such, we hypothesize that *interaction type* of psychological assessment would influence the assessment credibility **(H1)**.

Previous studies show that adding interactive elements (e.g., interactive probing and interactive feedback) to the questionnaire could improve the response quality for the follow-up open-ended questions [51, 61]. Compared with the form-based questionnaire, the conversation-based survey behaves as a virtual interviewer and intrinsically enriches interactivity through conversation, enhancing the response quality [14] and enjoyment [62]. Therefore, we hypothesize that the conversation-based psychological assessment would lead to higher enjoyment **(H2)** and higher response quality in open-ended questions **(H3)**.

2.3.2. Questionnaire Length

Numerous studies have investigated the effects of questionnaire length on a variety of indicators of a questionnaire, such as participation rates [53], dropout rates [54, 55], and response quality [25, 24]. Although longer questionnaires may discourage initial participation due to a higher response burden, no empirical evidence indicates “shorter is better” [63]. The short questionnaires are often criticized due to lower reliability [63]. As such, we hypothesize that the shorter questionnaire would negatively influence assessment credibility **(H4)**. Moreover, participating in a psychological assessment can enhance self-awareness [64], and a longer assessment requires users to spend more time reflecting on their mental status, which may increase mental health awareness. Thus, we hypothesize that a longer questionnaire could lead to a higher self-awareness of loneliness in our study **(H5)**.

According to a meta-analysis of response rates in web surveys [65], the length is not always associated with response rates. Nevertheless, adopting a longer questionnaire generally tends to decrease the response rate and cause a higher dropout rate [54, 23]. However, the quality of the responses does not necessarily deteriorate with a lengthy questionnaire as long as participants’ motivation can be maintained [25].

2.4. Closed-EQs versus Open-EQs

The *closed-ended questions* (Closed-EQs) and *open-ended questions* (Open-EQs) are two major types of questions in web surveys. Closed-EQs are more effective for gathering quantitative data [66], and Open-EQs perform better at measuring knowledge and obtaining more reliable and in-depth information [67, 16]. However, Open-EQs may increase the burden of the respondents [68] and the non-response rate due to more required cognitive efforts [69, 70]. Prior work showed the correlation between the responses to Closed-EQs and those to Open-EQs in web surveys for job satisfaction and user experience of e-service websites. Precisely, the dissatisfied employees, as measured via Closed-EQs about job satisfaction, were more likely to provide negative responses to Open-EQs [20] and disclose more content of negative feelings in Open-EQs [19]. Likewise, users with negative experiences of the e-service measured by Likert scale questions (a kind of Closed-EQs) tended to respond more to the comment-specific Open-EQs than those with positive experiences [18].

A mental health survey chatbot may ask users to answer Closed-EQs for a psychological assessment and Open-EQs for additional or detailed information regarding the assessment results. However, it is unclear how the psychological assessment design could influence user responses to Open-EQs in a survey chatbot for mental health. Previous studies have mainly revealed the relationship between Closed-EQs and Open-EQs based on user responses [18, 20, 19], while our work aims to investigate how the design aspects of Closed-EQs (i.e., interaction style and questionnaire length) influence users' responses to Open-EQs for collecting more in-depth data about mental health.

2.5. Perceptions of Mental Health Survey

Our study measures user perceptions of the mental health survey in terms of assessment credibility, self-awareness, and enjoyment.

2.5.1. Assessment Credibility

The users' perception of the psychological assessment results [71] (named *assessment credibility* in this work) is crucial as it could affect their health-related behaviors and decisions [72, 73]. Broadly speaking, the psychological assessment result is a type of health information. Previous studies have revealed several factors that could influence the perceived credibility of online health information, including source expertise [74, 75, 76] (i.e., the rating of the source), website design (e.g., layout, interactivity, visual design) [77, 76], the language used online [75], and ease of use [77].

2.5.2. Self-Awareness

Self-awareness refers to being conscious of users' own feelings, thoughts, beliefs, and behaviors, which is key to effective counseling and psychotherapy [78]. In the context of mental health, self-awareness is more about emotional self-awareness that can be gauged from four aspects: identifying emotions, empathy, managing emotions, and social skills [79]. Psychological assessment provides users with early problem detection and feedback, which in turn increases their self-awareness and general knowledge [64]. Thus, the design of these assessments is fundamental in fostering users' self-awareness regarding their mental health status.

2.5.3. Enjoyment

Enjoyment is a hedonic experience with which users deeply engage in an enjoyable activity [80]. Lin et al. [81] proposed a scale to measure enjoyment of the web experience based on three dimensions: engagement, positive affect, and fulfillment. Several studies have demonstrated the positive effects of chatbots on the effectiveness of surveys [62, 82] and the persuasion of health insurance recommendations [83], which are mediated by perceived enjoyment. Furthermore, enabling chatbot self-disclosure [45] or anthropomorphic cues [84, 85] can improve enjoyment, in turn promoting behavioral intentions (e.g., intention to use).

2.6. Evaluation of User Responses to Open-EQs

The main goal of asking Open-EQs is to collect richer data logically concerning response quality and self-disclosure [86]. Previous studies on survey chatbots evaluate user responses to Open-EQs mainly from response quality and the degree of self-disclosure [87, 21].

2.6.1. Response Quality

Compared to the responses to Closed-EQs, the responses to Open-EQs are free-form answers in an open text format, the quality of which can be gauged by some objective metrics such as response length, number of themes, response time, and item non-response [88]. For the Open-EQs in a chatbot, researchers employ Gricean Maxims (i.e., informativeness, specificity, relevance, and clarity) [21], readability [89], and sentiment intensity [90] to measure response quality.

2.6.2. Self-Disclosure

As an indicator of user engagement in chatbots, self-disclosure measures to what extent users would like to share their personal information, thoughts, and feelings [91], which is particularly important for the chatbot to understand the users' mental status [46]. Various self-reported instruments, such as Jourard Self-Disclosure Questionnaire (JSDQ) [92], Distress Disclosure Index (DDI) [93], and Self-Disclosure Index (SDI) [94], have been developed to measure self-disclosure by asking participants to rate their tendency to disclose information about their attitudes, opinions, and feelings on a Likert scale. Besides, the self-disclosure can also be rated by assessors from breadth (i.e., the range of discussed topics) and depth (i.e., the level of details discussed for a specific topic) [95]. Our study adopts both *subjective* and *objective* measurements to gauge self-disclosure in the user responses to Open-EQs. As the level of self-awareness is found to be positively related to self-disclosure during computer-mediated communication [96], we, therefore, hypothesize that users' self-disclosure is positively associated with self-awareness **(H6)**. Additionally, the credibility of health information could influence the self-disclosure of personal health information [97, 98]. As such, we hypothesize that a higher level of assessment credibility would lead to a higher degree of self-disclosure **(H7)** for Open-EQs.

3. Method

We employed a mixed method of qualitative and quantitative approaches to study how two design features of the psychological assessment (i.e., interaction style and questionnaire length) influence user perceptions of the assessment and user responses to Open-EQs.

3.1. Study Background

To address our raised research questions in a real-world setting of mental health service, we designed and developed a chatbot (called Percy) to help college students cope with loneliness during COVID-19 in collaboration with the Counseling and Development Center (CDC) of Hong Kong Baptist University (HKBU) that provides free and confidential counseling to students as well as consultation and referral services for staff. Participants were recruited through email invitations sent by the CDC of the university. We took precautions to minimize potential biases and priming effects by providing clear instructions and ensuring participants understood the purpose of the study without explicitly influencing their responses toward loneliness. Percy bot has three distinct functions: 1) psychological assessment of loneliness and overall mood [Figures 1(a-d)], 2) asking Open-EQ to get additional information about the feeling of loneliness [Figure 1(e)], and 3) offering some practical suggestions for managing loneliness [Figure 1(f)], for example, "*Call a friend or join an online group.*"

3.2. Participants

The study targets college students who experience loneliness during the COVID-19 pandemic. The Research Ethics Committee of Hong Kong Baptist University granted ethics [human (non-clinical)] clearance approval for this study. We recruited 330 participants using mailing lists and public bulletin boards for three weeks. As a result, 266 participants successfully finished the entire study. To ensure the quality of data, we filtered participants by four criteria: 1) the detected outliers (N=14) having extraordinarily long or short completion time based on the interquartile range (IQR), 2) the participants (N=10) who failed in two attention check questions, 3) the participants (N=7) who gave the meaningless responses (e.g., "nono" and "xxx") to all the Open-EQs, 4) the participants (N=22) who gave the same answers to all the questions asked in the post-study. Finally, we kept 213 valid participants for further analyses. Among those 213 valid participants, 80.28% of them (N=171) are female (because HKBU has a 1.7 : 1 ratio of female students to male students [3]), 89.67% of them (N=191) are 18 to 25 years old, 7.98% of them (N=17) are aged 25 to 30, and 2.35% (N=5) are older than 30. In addition, 78.87% of participants (N=168) are Hong Kong locals, and the rest are international students. To thank participants for supporting our research, 30 participants who completed the study were drawn to receive a supermarket coupon valued at 200 HKD (≈25.7 USD).

3.3. Design Manipulations

3.3.1. Manipulation of Interaction Style

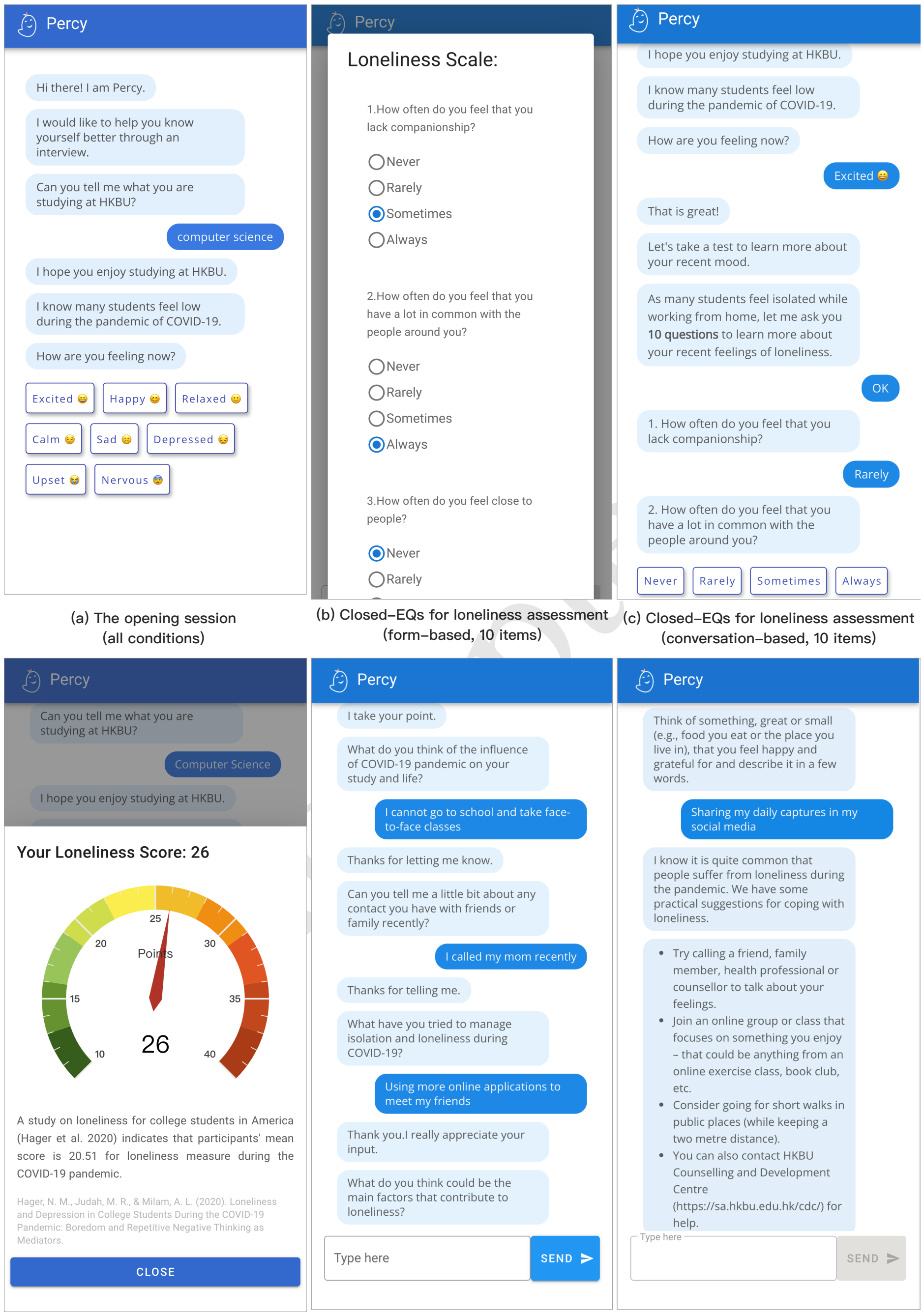
We offered two interaction styles for answering the questions in the psychological assessment: *form-based* and *conversation-based*. The choice of the two alternative interaction styles for the psychological assessment is based on reviewing the user interface design guidelines of several major conversational platforms such as Messenger[4] and WhatsApp[5]. For example, the form-based interaction is proposed based on the Webview in Messenger.

Form-based. The Percy bot offered an alternative way to present the questions of a psychological assessment in which all questions are embedded in a web form (see Figure 1(b)). We think the form-based interaction could increase psychological assessment efficiency while maintaining the interactivity of assessing their mental health in the chatbot.

[3] `https://intl.hkbu.edu.hk/student-exchange/incoming-students/why-hkbu/fast-facts`

[4] `https://developers.facebook.com/docs/messenger-platform`

[5] `https://www.facebook.com/brand/resources/whatsapp/user-interface`



(a) The opening session (all conditions)

(b) Closed–EQs for loneliness assessment (form–based, 10 items)

(c) Closed–EQs for loneliness assessment (conversation–based, 10 items)

(d) The assessment score of loneliness (based on 10–items Closed–EQs)

(e) Open–EQs asked after completing the Closed–EQs of assessment

(f) Practical suggestions for managing loneliness

Figure 1: Screenshots of Percy bot: (a) the opening session of conversation and mood recording, (b) the loneliness assessment with the web form, (c) the loneliness assessment in the conversation, (d) the result of loneliness assessment, (e) Open-EQ for getting additional information about the feeling of loneliness, and (f) practical suggestions for coping with loneliness.

Conversation-based. In this condition, all the loneliness psychological assessment questions were presented in the conversational style. Users can answer a question by clicking one of the buttons under the dialog in conversation that contains, for instance, selecting one from four options: "Never", "Rarely", "Sometimes", and "Always" (see Figure 1(c)). The transformation from a web survey to a conversational survey could improve response quality and user engagement [14, 62].

3.3.2. Manipulation of Questionnaire Length

The longer questionnaire can result in a "straight-line" response pattern, which means more identical answers to most Closed-EQ [25]. Thus, we think the questionnaire length could influence users' patience and carefulness in the psychological assessment. Moreover, the increased response burden caused by a long questionnaire may influence response quality and response length for Open-EQs.

In this study, our chatbot specializes in surveying university students' loneliness during the pandemic of COVID-19. UCLA loneliness scale is the most widely used instrument for assessing loneliness [17], and it has three validated length versions, including three items, ten items, and twenty items, respectively [99, 17]. Based on the three versions, we determined three questionnaire lengths: short (three items), middle (ten items), and long (twenty items). The questions in the short version are measured on a three-point scale (1 = Hardly Ever; 2 = Some of the Time; 3 = Often) [99], while the questions in the middle and long versions are rated on a four-point scale (1 = Never; 2 = Rarely; 3 = Sometimes; 4 = Always) [17].

3.4. User Study Design and Procedure

Based on our two independent variables, *interaction style* and *questionnaire length*, we designed a 2 (interaction style: form-based vs. conversation-based) $\times$ 3 (questionnaire length: short vs. middle vs. long) between-subjects study. Figure 2 shows an overview of the study design, including the following three major phases:

Pre-study. First, we asked all participants to sign a consent form and read an information page describing Percy's main features and explaining the steps they should follow to finish the study. After that, we asked participants to answer three questions about their demographics, including age, gender, and nationality.

Moreover, we asked participants to indicate their current mood from eight options based on two dimensions of core-affect [100], including excited, happy, relaxed, calm, sad, depressed, upset, and nervous (Figure 1(a)).

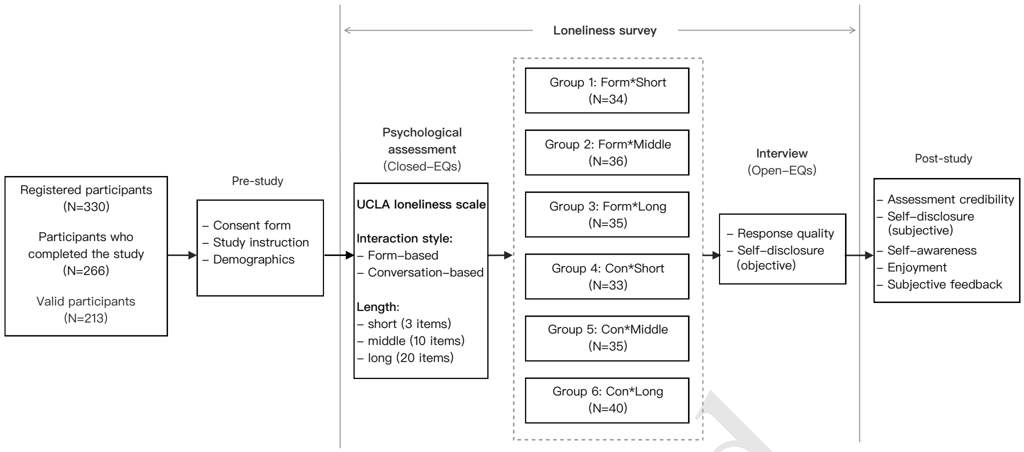


Figure 2: User study design and procedure.

Loneliness survey. The loneliness survey contains a psychological assessment (measured by Closed-EQs) and an interview (measured by Open-EQs). The psychological assessment has six variants combining two design manipulations: *interaction style* and *questionnaire length*. Following the between-subjects design, we randomly assigned participants to one of six conditions. When users finished the psychological assessment, a result page popped up, showing a loneliness score, a semicircle meter with color gradients for the score level, and an explanation with a reference for the score (Figure 1(d)). The participants were then guided to the interview session after closing the result page. During the interview, the chatbot asked seven Open-EQ (see Table A.4 in Appendix A) to understand the participants' feelings of loneliness during COVID-19 deeply. As the chatbot's responses may likely influence how users chat with it [21], our chatbot only generated some general responses to users' answers to avoid such interference. These responses vary and depend on the content of users' answers, for example, *"Thank you. I appreciate your input."* or *"Thank you for your thoughtful input."* are possible responses for the user answers of rich content, e.g., *"I wish to be around my family more often where I can be myself more. I also think exercising regularly can help."*, while *"Got it."* or *"I understand!"* are for simple and brief user answers, e.g., *"It's fine."* or *"nothing"*.

Post-study. Participants were required to complete a questionnaire containing sixteen five-point Likert scale questions (Table 1) to indicate their perceived assessment credibility, self-awareness, enjoyment, and self-disclosure. In addition, we asked participants to answer five Open-EQs (see Table B.5 in Appendix B) to understand their in-depth opinions on Percy.

3.5. Measurement and Analysis

This study measured users' perceptions of the loneliness survey based on assessment credibility, self-awareness, and enjoyment. Moreover, we adopted several metrics for response quality and subjective and objective measures for self-disclosure in user responses to Open-EQs.

Table 1: Post-Study Questionnaire for Measuring User Perceptions of the Survey and Self-Disclosure

Construct	Item	Loading
Assessment Credibility (Cronbach alpha: 0.894; AVE: 0.741)		
	I am convinced that the score can indicate my feelings of loneliness.	0.709
	I am confident I will trust my loneliness score.	0.770
	The loneliness score calculated by the Percy bot can be trusted	0.674
Self-Awareness (Cronbach alpha: 0.818; AVE: 0.607)		
	I have insight into myself.	
	I recognize the stress and worry in my current life.	0.696
	I understand myself well.	
	I generally feel positive about self-awareness.	0.581
	The Percy bot made me aware of my loneliness.	0.754
Enjoyment (Cronbach alpha: 0.841; AVE: 0.649)		
	I enjoy talking with the Percy bot.	0.716
	I feel enjoyable when I converse with the Percy bot.	0.798
	I would like to answer survey questions with the Percy bot.	0.612
Self-Disclosure (subjective) (Cronbach alpha: 0.758; AVE: 0.610)		
	I think I have told my real feelings to the Percy bot.	0.605
	I think I have provided sufficient information to the Percy bot.	0.578
	The design of the interview Percy bot made me think longer about my responses compared to traditional surveys.	
	If time allows, I would like to spend more time elaborating my responses to let the Percy bot understand me better.	
	I am not willing to reveal my feelings to the Percy bot. (reversed)	

Note: The items marked in gray were dropped due to a poor loading value (< 0.5) or high cross-loading value (> 12) measured by modification index [101].

3.5.1. Perceptions of Loneliness Survey

Perceptions of the loneliness survey refer to participants' feelings and attitudes towards the loneliness assessment (Closed-EQs) and the interview (Open-EQs). We employed a set of questions (see Table 1) to measure three constructs: assessment credibility, self-awareness, and enjoyment. All these questions were measured on a five-point Likert scale. We run a confirmatory factor analysis (CFA) to establish the validity of these question items. Commonly accepted cutoff values for convergent validity are 0.7 for Cronbach's alpha, 0.5 for average variance extracted (AVE) [102], and 0.5 for factor loading.

- *Assessment credibility.* It measures to what extent the psychological assessment result can be trusted and believed. According to Hilligoss and Rieh's credibility framework consisting of three levels of credibility judgments: construct, heuristics, and interaction [103], We composed three questions to measure participants' perceived credibility of their loneliness assessment (Cronbach alpha: 0.894; AVE: 0.741).

- *Self-awareness.* Self-awareness is the participant's ability to know and understand their feelings and behaviors. We measured self-awareness based on the three validated questions of a Self-Awareness Outcomes Questionnaire (SAOQ) [104] (Cronbach alpha: 0.818; AVE: 0.607).

- *Enjoyment.* It gauges how much the participants enjoyed chatting with Percy. We used three validated questions from a questionnaire for evaluating recommendations in a mental health app [105] to measure enjoyment (Cronbach alpha: 0.841; AVE: 0.649).

3.5.2. Response Quality

In this study, we did not measure the response quality of Closed-EQs using methods such as differentiation response index (i.e., satisficing behavior of choosing the same response every time) [106] because these metrics are usually applied to assessing whether participants are serious and attentive for answering the questions in general surveys such as internet usage behavior [14] and course satisfaction [62]. In our opinion, the motivation for completing a mental health survey differs from answering a general survey. The participants are more motivated by a need to understand their mental health status more accurately. Moreover, choosing the same response to all the questions in a short psychological assessment (e.g., the short loneliness assessment with five Closed-EQs) does not necessarily mean satisfying behavior.

For Open-EQs, we measured the response quality based on Gricean Maxims theory [107] that has often been used to evaluate the quality of users' responses in chatbots [87, 21]. Gricean Maxims was developed based on the cooperative principle for enabling effective conversational communication by concretely considering four aspects: quantity, quality, relevance, and manner [108]. According to the definition of Gricean Maxims, the aspect of "quality" refers to being truthful in communication. Due to the general difficulty in assessing the truthfulness of user responses [21], we did not measure this aspect. In our study, we concretely adopted four quality metrics (i.e., informativeness, specificity, relevance, clarity) used to evaluate user responses to Open-EQs in a chatbot [21], which were proposed based on three Gricean Maxims aspects: quantity, relevance, and manner (see Table 2). We measured these metrics based on user responses to all Open-EQ asked by our Percy bot.

Table 2: Quality Metrics Defined Based on Gricean Maxims [21]

Gricean Maxims	**Definition**	**Quality Metric**	**Definition**
Quantity	One should be as informative as possible.	*Informativeness*	A participant's response should be as informative as possible.
		Specificity	A participant's response should give as much information as needed.
Relevance	One should provide relevant information.	*Relevance*	A participant's response should be relevant to a question.
Manner	One should communicate in a clear and orderly manner.	*Clarity*	A participant's response should be clear.

- *Informativeness.* Per the maxim of quantity, the communication should be as informative as possible. The measure of informativeness in users' responses based on Formula (1) [21] that calculates the sum of a word's surprisal based on the inverse of its occurrence frequency in four major English corpora, including British National Corpus [109], the Brown Corpus [110], Webtext [6], and the NPS Chat Corpus [111].

[6] https://github.com/teropa/nlp/tree/master/resources/corpora/webtext

$$I(Response) = \sum log_2 \frac{1}{F(word_n)} \quad (1)$$

- *Response quality index.* We measured the overall response quality by response quality index (RQI) [21] that combines three quality metrics: specificity, relevance, and clarity, as shown in Formula (2) and respectively defined in Table 2. The measures of the three quality metrics follow a manual assessment method, and we defined three levels (0,1,2) for each metric. In total, we collected 1,491 text responses from 213 participants. We followed a standard coding protocol to code each response. First, we randomly selected 10% of responses and then asked two researchers to finish the coding independently. After that, they discussed the differences in coding, and a third researcher was involved in voting for the irreconcilable differences. The coding criteria became more consistent after the discussion. Finally, they finished coding for the rest of the responses. The Cohen's kappa of each set of coding (Specificity: κ=0.73, Relevance: κ=0.81, Clarity: κ=0.89) indicates good inter-rater reliability of the coded items [7].

$$RQI = \Sigma_{n=1}^{N} specificity[i] * relevance[i] * clarity[i] \quad (2)$$
(N is the number of responses in a completed assessment)

Table 3 shows some examples of our coded responses. *Specificity* refers to the level of details the response provides, and a specific response should convey meaningful insights (0 – generic description only, 1 – specific concepts, and 2 – specific concepts with detailed examples). *Relevance* measures to which extent the answer is relevant to the question asked during the interview (0 – irrelevant, 1 – somewhat relevant, and 2 – relevant). *Clarity* is measured based on the human effort of understanding the text (0 – illegible text, 1 – incomplete sentences, and 2 – clearly articulated response).

3.5.3. Self-Disclosure

Self-disclosure involves sharing personal thoughts, feelings, or experiences about oneself with others [113]. The quality of user responses to Open-EQs in a survey is linked to the extent of self-disclosure [86], signifying the extent to which

[7] Slight: 0.0-0.2; Fair: 0.21-0.4; Moderate: 0.41-0.6; Substantial: 0.61-0.8; Almost Perfect: 0.81-1 [112].

Table 3: Examples of Coded Responses to the Open-Ended Question Open-EQ7 ("Think of something that you feel happy and grateful for, great or small (e.g., *the food you eat or the place you live in*).")

Response Example	**Rating**
"my family, including my father, even though he had passed away. Also, my husband. All about love; I know they love me even though I don't know how to express the gratitude."	Specificity:2, Relevance:2, Clarity:2, Self-disclosure:2
"Money"	Specificity:2, Relevance:1, Clarity:0, Self-disclosure:0
"Listening to my favorite music and watching my favorite reality show ."	Specificity:2, Relevance:2, Clarity:1, Self-disclosure:1
"Everything will be fine."	Specificity:1, Relevance:2, Clarity:0, Self-disclosure:0

users are willing to share information with the chatbot. In Open-EQs, we assessed self-disclosure based on users' subjective feelings and objective metrics of user responses, such as the breadth and depth of content.

- *Self-disclosure (subjective).* It assesses participants' subjective perspectives on sharing their feelings and thoughts about loneliness. The questions for measuring subjective self-disclosure, as depicted in Table 1, have been adapted from those used to evaluate user responses in a survey chatbot [62] (Cronbach's alpha: 0.758; AVE: 0.610).

- *Self-disclosure (objective).* It gauged the extent to which participants shared their personal feelings and thoughts with the chatbot. We manually evaluated the level of self-disclosure based on the breadth and depth of topics conveyed in user responses to the seven Open-EQs (0 – a brief description with no specific topic, 1 – a brief description with a specific topic, and 2 – a detailed description with one specific topic / a description with multiple topics) [91]. The self-disclosure coding demonstrated substantial inter-rater reliability, as evidenced by Cohen's kappa score of 0.69. As illustrated in the example (the first example in Table 3), higher levels of self-disclosure may encompass more detailed and private topics.

3.6. Interaction Behavior

We also recorded response length for Open-EQs and engagement duration to understand better how much users would like to interact with the chatbot.

- *Response length.* Response length was counted by the number of words in each participant's responses to all seven Open-EQs during the interview. The response length is usually proportional to the engagement duration.

- *Engagement duration.* Engagement duration measured the time a participant spent answering all the Open-EQs in the interview session of the loneliness survey. A longer engagement duration could mean the participant invests more effort in thinking and answering the Open-EQ.

4. Results

This section presents the main results related to each research question. For the convenience of illustration, we use an expression of **interaction*length** to denote each experimental condition in the remaining parts of this manuscript. In this expression, interaction can be "Con" or "Form", respectively standing for *conversation-based* and *form-based*, and length can be "Short", "Middle", or "Long". For example, Con*Middle refers to the condition where participants assessed their loneliness by completing the middle-length UCLA loneliness scale (ten items) through conversation-based interaction for Closed-EQs.

To investigate two design factors (i.e., interaction style and questionnaire length), we employed a 2x3 factorial design in our study. Additionally, we need to run multiple regression analyses to test our research hypotheses. To achieve this, we have opted to use structural equation modeling (SEM) to analyze our results, given its capacity to evaluate multivariate causal relationships simultaneously within a statistical estimation procedure [114]. Table C.6 in Appendix C presents the descriptive statistics of the dependent variables (DVs) for six experimental conditions derived from a 2x3 factorial design.

4.1. Structural Equation Modeling

We use *lavaan*, [8] an R package to build our SEM model. Some dependent variables (DVs), such as informativeness, engagement duration, and response length, were measured differently from the five-point Likert scale for measuring the DVs related to user perceptions, resulting in much larger values. Therefore, we normalized the values of these dependent variables by using the *scale*() function in R, which scales the data based on the mean value and the standard deviation. In addition, as our data do not conform to the normal distribution, we chose a more

[8] `https://lavaan.ugent.be/`

robust estimator, "MLR," in our SEM analysis. The sample size of our study meets a CFA/SEM rule of thumb that 10:1 is the recommended ratio of subjects to observable variables (N:q) [115] and the recommended sufficient sample size (N = 200) for structural equation modeling [116, 117]. Following the procedure of trimming non-significant paths in SEM model [118], we obtain our resulting model (see Figure 3) showing a good fit [9]: $\chi^2(149) = 209.323$, p=.003 [10]; root mean squared error of approximation (RMSEA) = 0.044; 90% CI: [0.029, 0.057]; Comparative Fit Index (CFI) = 0.969; Turker-Lewis Index (TLI) = 0.963. In addition, we utilized the R package, *semPower*, [11] to execute a post-hoc power analysis for our obtained model. The analysis revealed a high power level (power > .98) with a sample size of N = 213 to identify misspecifications of a model (involving df = 149 degrees of freedom) corresponding to RMSEA ≥ .05 at an alpha error level of .05.

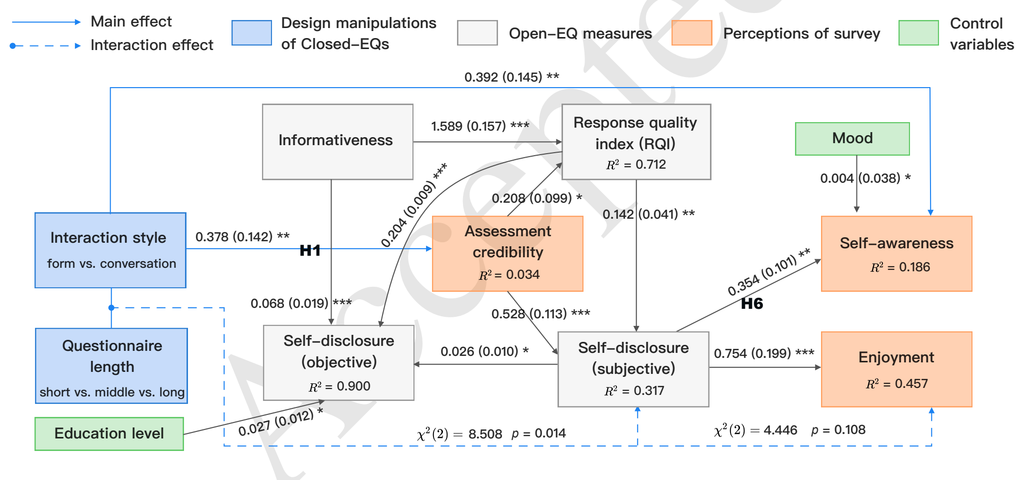


Figure 3: The structural equation model for our user study's data. Significance levels: *** p <.001, ** p <.01, * p <.05. The numbers on the edges refer to the β coefficient and standard error (in parentheses) of the causal relationship. R^2 is the proportion of variance explained by the model. Factors are scaled to have an SD of 1. The paths labeled with H1 and H6 indicate these two paths support hypotheses H1 and H6.

[9] Hu and Bentler [119] proposed cutoff values for several fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05, with the upper bound of its 90% CI below 0.10.

[10] A model should not have a non-significant χ^2, but this statistic is regarded as too sensitive [120].

[11] `https://github.com/moshagen/semPower`

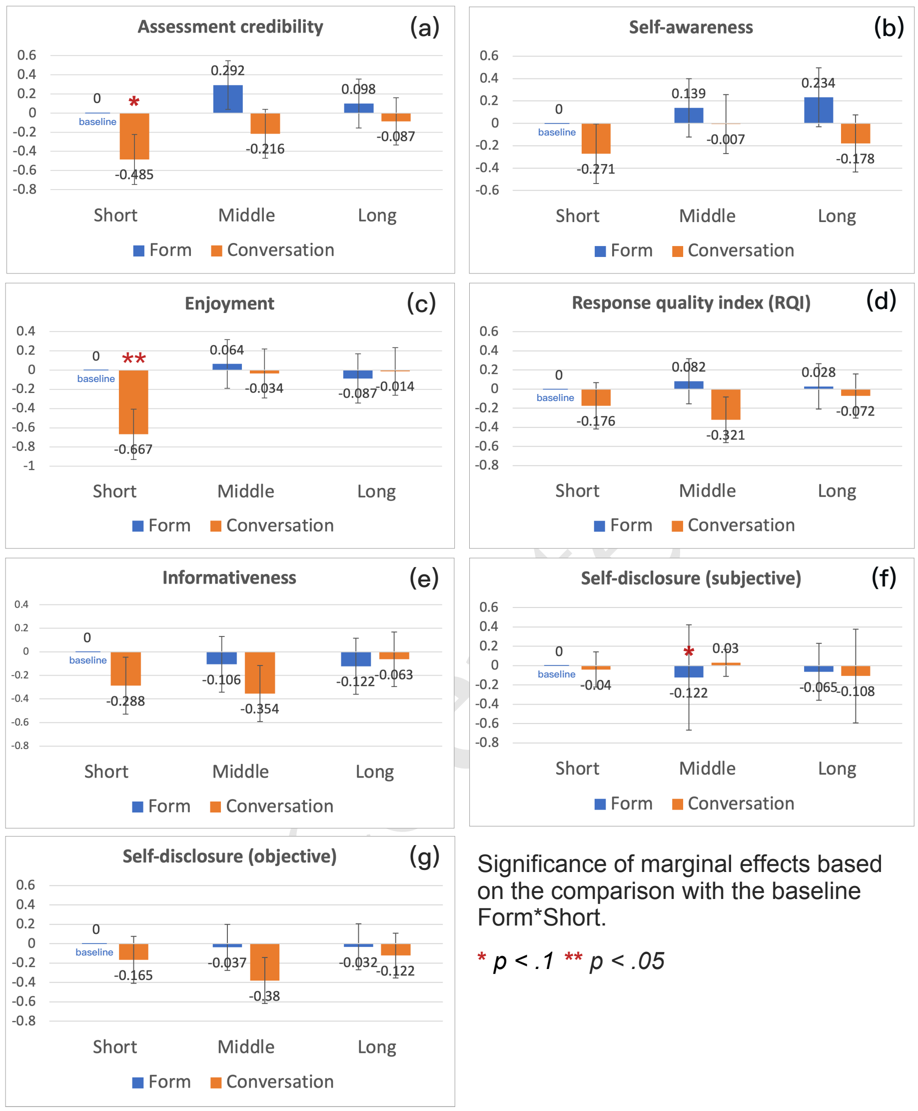


Figure 4: Marginal effects of interaction style and questionnaire length on different DVs. The effects of the baseline Form*Short are set to zero, and the y-axis is scaled by the sample standard deviation. Significance levels: **p <.05, * p <.1.

In addition, to understand how the values of a dependent variable (e.g., assessment credibility) change with variation of the independent variable (IV) (e.g., interaction style), we analyzed the marginal effects of the two IVs (i.e., interaction style and questionnaire length) on each DV, assuming other covariates to be fixed [121]. Figure 4 shows the marginal effects of dependent variables that are associated with significant main effects or interaction effects of two design factors. In order to effectively gauge or test our hypothesis, we also consider the potential influence of control variables (such as age, gender, education level, and mood) on the dependent variable. The findings indicate that education level significantly impacts self-disclosure, while mood significantly affects self-awareness.

4.2. The Effects of Interaction Style with Closed-EQs on Perceptions of the Survey (RQ1)

The SEM model (Figure 3) shows a direct positive effect of interaction style on assessment credibility (β = 0.378, p <.01). Moreover, as depicted in Figure 4(a), the conversation-based design appears to compromise user-perceived assessment credibility, particularly when combined with the short questionnaire. Con*Short was lower than the baseline with marginal significance (p <.1). Thus, we can accept the hypothesis **H1**: the form-based psychological assessment would lead to higher assessment credibility. Moreover, the model shows no other significant effects of interaction style on enjoyment and response quality in Open-EQs. Thus, we cannot accept the hypothesis **H2**: the conversation-based psychological assessment leads to higher enjoyment, and the hypothesis **H3**: the conversation-based psychological assessment leads to higher response quality in Open-EQs. The marginal effects on enjoyment (Figure 4(c)) indicate that combining conversation-based interaction and a short questionnaire could lower enjoyment, and Con*Short is significantly lower than the baseline in terms of enjoyment (p <.05). In addition to testing our hypothesized effects, the model shows a significant effect of interaction style on self-awareness (β = 0.392, p <.01). The marginal effects on self-awareness (Figure 4(b)) show that form-based interaction leads to higher self-awareness than conversation-based interaction regardless of the questionnaire length.

4.3. The Effects of Questionnaire Length on Perceptions of the Survey (RQ2)

Manipulating questionnaire length does not directly affect any investigated measures for users' perceptions of the survey. Thus, we could not accept the hypothesis **H4**: a shorter questionnaire leads to lower assessment credibility, and the hypothesis **H5**: a longer questionnaire leads to higher self-awareness. Even though not statistically significant, users seem to perceive higher assessment credibility with the form-based design when completing a middle questionnaire (refer to Figure 4(a)), and they attain increased self-awareness by completing a longer questionnaire (as seen in Figure 4(b)). Furthermore, we find an interaction effect of interaction style and questionnaire length on enjoyment, which is marginally significant, $\chi^2(2) = 4.446$, $p = .108$. In other words, the effects of questionnaire length on enjoyment depend on the interaction style. Specifically, the distinction between the short questionnaire and questionnaires of other lengths is more pronounced with conversation-based interaction than with form-based interaction (see Figure 4(c)).

4.4. The Effects of Interaction Style with Closed-EQs on User Responses to Open-EQs (RQ3)

The SEM model (Figure 3) does not show any direct effect of interaction style on response quality and self-disclosure measures. Despite no significant direct main effects of interaction style on response quality, the form-based design could positively influence self-disclosure (subjective and objective) and RQI through assessment credibility. The assessment credibility positively influences self-disclosure (subjective) (β = 0.528, p <.001) and RQI (β = 0.208, p <.05), which in turn positively influences self-disclosure (objective). Thus, the significant effects of assessment credibility on self-disclosure (subjective) and self-disclosure (objective) allow us to accept the hypothesis **H7**: higher credibility leads to more self-disclosure in Open-EQs.

Specifically, the significant paths (*P1*: Interaction style → Assessment credibility → Self-disclosure (subjective) → Self-disclosure (objective)) and (*P2*: Interaction style → Assessment credibility → RQI → Self-disclosure (objective)) indicate a *mediating role* of assessment credibility in the effects of interaction style on self-disclosure (objective) in Open-EQs. Figure 4(g) shows that regardless of the questionnaire length, conversation-based interaction results in lower levels of self-disclosure (objective) compared to form-based interaction. However, the total indirect effect of assessment credibility on self-disclosure (objective) is minimal (β = 0.057).

4.5. The Effects of Questionnaire Length on User Responses to Open-EQs (RQ4)

The model does not show any main effects of questionnaire length on response quality. The marginal effects of questionnaire length on RQI and informativeness illustrate the non-significant difference caused by the manipulation of questionnaire length (see Figure 4(d and e)). Compared with the baseline condition (Form*Short), the short and middle questionnaires lead to lower response quality with the conversation-based design.

Despite no main effect of questionnaire length on self-disclosure measures, we find a significant interaction effect of interaction style and questionnaire length on self-disclosure (subjective), $\chi^2(2) = 8.508$, $p < .05$, indicating that the effect of questionnaire length on self-disclosure (subjective) depends on interaction style. For instance, the marginal effect on subjective self-disclosure (Figure 4(f)) indicates that the middle questionnaire results in the highest subjective self-disclosure, with marginal significance ($p < .01$) when combined with conversation-based interaction, whereas it leads to the lowest subjective self-disclosure when combined with form-based interaction.

4.6. Relations Between User Responses to Open-EQs and Perceptions of the Survey

The model also reveals the relationships between the perceptions of the survey (i.e., enjoyment and self-awareness) and user responses to Open-EQs. Specifically, the significant path (P3: Informativeness → RQI → Self-disclosure (subjective) → Self-Awareness & Enjoyment) confirms the mediated effects of informativeness and response quality on self-awareness and enjoyment. As self-disclosure (subjective) positively influences self-awareness ($\beta = 0.354$, $p < .01$), we could accept the hypothesis **H6:** higher self-disclosure is positively associated with self-awareness. Interestingly, self-disclosure (subjective) has a strong positive effect on enjoyment ($\beta = 0.754$, $p < .001$), indicating that participants who are willing to disclose their personal feelings and experiences are more likely to perceive enjoyment while interacting with the survey chatbot. Moreover, the significant path (P4: Assessment credibility → Self-disclosure (subjective) → Self-Awareness & Enjoyment) suggests that participants who perceive higher assessment credibility tend to disclose their feelings and thoughts about loneliness with the chatbot and then perceive higher self-awareness and enjoyment.

4.7. Interaction Behavior

We recorded the number of words in each participant's responses to all Open-EQs (response length) and the total time they spent answering them (engagement duration). Design manipulations do not directly affect response length and engagement duration. Nevertheless, the conversation-based interaction leads to shorter responses than the form-based interaction, and the condition of Form*Middle has the longest response on average (M=60.9 words, SD=44.6). Furthermore, the questionnaire length positively influences engagement duration when adopting the conversation-based interaction, and the condition of Form*Middle has the longest engagement duration (M=339.8 seconds, SD=277.7).

4.8. Subjective Feedback

To better understand participants' subjective experiences of two design manipulations in our survey chatbot, we performed a thematic analysis [122] based on participants' responses to the five Open-EQs in the post-study (Table B.5). Two authors independently finished half of the responses and addressed the conflicts in coding through additional discussion, resulting in an almost perfect inter-rater agreement among coding tested by Cohen's kappa ($\kappa = 0.85$) [12]. One author finished coding the remaining responses and discussed them with another author to reach a consensus on the codes.

The Length of Questionnaire. Using a short questionnaire could potentially diminish the credibility of the assessment. Although the questionnaire length does not significantly influence assessment credibility according to the quantitative analysis, a short questionnaire seems to decrease users' perceived assessment credibility. Certain participants trying with the short version of the assessment believed that incorporating more question items could enhance the credibility of the test, as two participants noted,

> *"I think there could be more questions to indicate my loneliness score better."* (P7, Form*Short)

> *"I don't think people's loneliness can be scored when people just answer three questions."* (P170, Con*Short)

Interaction Style of Psychological Assessment. Moreover, compared with the form-based interaction, the conversation-based interaction offers a more casual way for users to answer the questions measured by the Likert scale. However, it may also make the questionnaire perceived as less formal, aligning with the result of quantitative analysis. One participant called,

> *"It is just like chatting. But I don't really agree with the score, and it may need an adjustment to have more options. Maybe 0 to 10."* (P40, Con*Middle)

It seems that presenting the questions of a psychological assessment via the conversational style decreases the questionnaire's formality [60], which in turn influences users' perceived assessment credibility. However, some participants

[12] Slight: 0.0-0.2; Fair: 0.21-0.4; Moderate: 0.41-0.6; Substantial: 0.61-0.8; Almost Perfect: 0.81-1 [112].

doubted the assessment's credibility because of the ambiguous measurement standard for loneliness; for example,

> *"...these are some general questions, cannot be sure if the score is trustworthy cause people have different standards."* (P206, Con*Middle)

Additionally, a few participants also complained about the increased interaction time caused by the conversation. For example, one participant stated,

> *"Filling in an online form can be boring if there are too many questions. Chatting with the Percy bot is interesting, at least with more interaction. But chatting with a bot can be time-consuming."* (P137, Form*Middle)

Psychological Assessment Result. The assessment score is key to self-awareness. Many participants claimed that they became more aware of their loneliness status by finishing the psychological assessment. One participant noted,

> *"I think the questions asked were relevant for calculating the loneliness score. I am aware of what my feelings are during the pandemic."* (P108, Form*Middle)

Some participants thought the reference on the result page (see Figure 2(d)) showing the mean score of others who completed this loneliness assessment helped them better understand their loneliness status.

> *"Comparing to the mean score, I know more about my status among people."* (P137, Form*Short)

5. Discussion

Prior research has highlighted the benefits of using a survey chatbot as compared to a conventional survey delivered through web forms. This study delves deeper into the refined design aspects of a survey chatbot within the scope of mental health. More specifically, we explore the impact of the interaction style and length of psychological assessments featuring Closed-EQs on the quality of responses to subsequent Open-EQs within a survey chatbot. Thus, the findings from this investigation are contextualized within a survey chatbot environment that presents both Closed-EQs and Open-EQs.

Before discussing the results of our study, we first briefly summarize our research findings based on quantitative and qualitative results.

1. **The interaction style of psychological assessment significantly affects the assessment credibility and self-awareness. The influenced assessment credibility could influence response quality and self-disclosure for Open-EQs.** The participants who completed the psychological assessment via the form-based interaction were more convinced by the assessment, thereby being more engaged in responding to the follow-up Open-EQs and being more aware of their feelings.

2. **The questionnaire length does not significantly impact the assessment credibility and user responses to Open-EQs.** Although there is an interaction effect between interaction style and questionnaire length on self-disclosure (subjective) and enjoyment, questionnaire length has no significant main effect on any dependent variables.

3. **The assessment credibility mediates the effects of psychological assessment design on users' responses to Open-EQs.** The psychological assessment design has *indirect* positive impacts on users' self-disclosure (objective) and response quality index (RQI) through the assessment credibility.

5.1. Psychological Assessment Design

The psychological assessment is vital for monitoring mental health status and delivering timely adaptive interventions in a mental health survey chatbot [123]. This is especially crucial when access to mental health services is limited, as seen during events like the COVID-19 pandemic [124]. With this in mind, our investigation focuses on how the design of the psychological assessment with Closed-EQs could impact users' perceptions of the assessment and their responses to Open-EQs in a survey chatbot.

5.1.1. Interaction Style of Closed-EQs

Our study investigated two interaction styles of psychological assessment with closed-ended questions in a survey chatbot: form-based and conversation-based. Previous studies have demonstrated the benefits of conversation-based design over form-based design for the entire survey in terms of response quality [14, 62, 21], without making a distinction between Closed-EQs and Open-EQs. However, we found that within a survey chatbot, the form-based interaction leads to higher assessment credibility with Closed-EQs, which in turn leads to higher response quality in Open-EQs. We argue that survey design for psychological assessments is different from surveying course satisfaction [62], gamers' opinions [21], and

Internet usage behavior [14] in previous studies. In contrast to traditional surveys, the psychological assessment is frequently succeeded by an assessment score or report, aiming to provide users with an understanding of their health status and encourage positive health behavior changes [50]. This process may lead participants to take the assessment questions more seriously, as inaccurate self-assessments could potentially impact mental health [125].

Despite the benefits of casual communication (e.g., more communicative [126], or a strong feeling of being involved [127]), formal communication has been proven to be associated with high information credibility [128]. Furthermore, a prior study showed that with a task-oriented chatbot, users are more likely to feel like performing a task in a natural, casual, informal conversation rather than in goal-directed settings [129]. Therefore, we speculate that the casual communication conveyed by the conversation-based design may decrease the users' perceived formality of assessment and weaken their perceived assessment credibility.

Moreover, our study shows that the conversation-based interaction significantly increases interaction time than the form-based one while adopting a long questionnaire (Figure 4(i)), which aligns the findings of a previous study on a survey chatbot with Closed-EQs [14]. Unlike the responses to Open-EQs, which could be diverse free-text inputs, the responses to Closed-EQs are based on predefined content, such as the Likert scale or multiple-choice. We think that the increased response time of the psychological assessment may imply a lower efficiency of assessment rather than higher user engagement. The conversational interaction may especially cause users' displeasure at the slow pace of completing a long questionnaire. Therefore, we wonder how we may make a trade-off between the advantages of the conversation-based design (e.g., natural interaction, less non-differentiation in a rating task, aka a "straight-line" response [14]) and its disadvantages (e.g., low efficiency). For example, one participant (P137, Form*Short) stated, "*Chatting with the Percy bot is quite interesting, at least with more interaction. However, chatting with a bot can be time-consuming.*" Thus, a form-based design could be more suitable for presenting a questionnaire in a chatbot because it maintains the formality and efficiency of the questionnaire and does not influence users' perceived interactivity of responding to the follow-up Open-EQs in the survey chatbot.

Therefore, we suggest **adopting a form-based design for the psychological assessment in a survey chatbot for mental health.** Although the conversation-based design has distinct advantages over the form-based design, such as interactive content [14], reciprocity [45], and human-like communication [44, 21], it also imposes more interaction time on users [14, 21]. More notably, the form-based design makes participants perceive higher assessment credibility than the conversation-based. Therefore, chatbot designers could embed a form-based psychological assessment into the chatbot before asking Open-EQs through conversation. This hybrid design may also combat the survey-taking fatigue in case the participants are expected to be more engaged in responding to the Open EQs [21]. On the one hand, users may feel they are still answering questions in the chatbot; on the other hand, they may focus more on questionnaire content, which is less tedious than following the humdrum conversation pattern to answer Closed-EQs.

5.1.2. The Length of Questionnaire with Closed-EQs

Information completeness is a major factor that influences the perceived credibility of health information [130]. The length of the questionnaire reflects how much information is collected for assessment, which could affect the completeness of the assessment information. Thus, we investigated how the questionnaire length influences the assessment credibility. However, we did not find a significant main effect of questionnaire length on users' perceived assessment credibility, probably because the participants did not perceive significantly different assessment results regarding information completeness with three different questionnaire lengths (short, middle, and long). Moreover, our results also indicate that questionnaire length does not have a significant main effect on the response quality and self-disclosure in Open-EQs, which echoes the findings of prior work that the response quality of Open-EQs is not associated with the survey length [63, 24]. Thus, keeping the assessment as short as possible is unnecessary, but the content (questions) of the psychological assessment should satisfy the users' assessment needs [63]. Additionally, the significant interaction effects of interaction style and questionnaire length on enjoyment and subjective self-disclosure in the follow-up Open-EQs suggest that the determination of questionnaire length might also depend on the questionnaire's interaction style. Therefore, we suggest that **designers may determine the questionnaire length based on user needs and the interaction style of the questionnaire.**

Moreover, according to a recent literature survey on the instruments used in the psychological assessment of mental health and health behavior [50], among 21 surveyed questionnaires (e.g., GAD-7 for anxiety [40], PHQ-7 for depression [39], PHQ-15 for physical symptoms [131]), the questionnaire length varies from 2 to 28 items, similar to the range used in our study. Consequently, our findings regarding the impact of questionnaire length could potentially be applied to scenarios utilizing other psychological assessments.

5.1.3. Assessment Credibility

Users' perceived credibility of health information significantly impacts their behavioral intention of using the health informatics service [132]. In our study, the structural equation model (Figure 3) demonstrates a mediating role of assessment credibility in the effects of the interaction style of psychological assessment on the metrics evaluating users' responses to Open-EQs. The users' perceived credibility of assessment is critical to the mental health survey, as it could influence user engagement in the activities at a later stage [132], for example, answering Open-EQs in a mental health survey.

Online health information can be categorized mainly into scientific and experiential information [133]. The results of the psychological assessment provided by the agent belong to scientific information, the credibility of which is mainly assessed based on reference credibility [133]. Thus, our psychological assessment result (score) page also shows an academic reference (Figure 1(d)) to justify the interpretation of the assessment score (Figure 1(d)). However, we wonder if participants could notice the study's reference and how much it may help them justify the result. Our qualitative results indicate that although we provide a descriptive explanation of the psychological assessment results based on a reference (Figure 1(d)), some participants still do not trust the assessment score due to the ambiguous measurement standard for loneliness, for example, *"...cannot sure if the score is trustworthy cause people have different standards." (P206, Con*Middle)* Therefore, the future design may allow users to ask for further explanations of the psychological assessment results through conversation. When addressing user inquiries about assessment results, the conversational explanation may be considered more convincing by users due to the persuasive potential of the chatbot [83].

In general, the credibility of information on the web can also be influenced by multiple aspects of the information medium, such as content format, design of user interface, and interactivity [59]. With the evolution of human-computer interaction, virtual agents' simulated human-human interaction is increasingly popular for mental health because of greater interactivity that supports therapeutic conversation [134]. However, should we deliver all the services in a mental health chatbot through conversation? For the psychological assessment, our study results suggest that the participants perceived higher assessment credibility with the form-based assessment questionnaire than with the conversation-based questionnaire. As most mental health surveys still adopt form-based questionnaires, the conversation-based interaction style probably does not conform to the participants' mental model of taking a psychological assessment.

5.2. User Responses to Open-EQs

We evaluated user responses to Open-EQs in our survey chatbot from multiple aspects, among which self-disclosure and response quality have more often been emphasized in the previous studies [44, 21, 45].

Self-disclosure refers to revealing personal and even sensitive information to others [135]. Prior work has identified its important role in building trust [136] and intimacy [137] for communication. In our study, users' subjective self-disclosure is satisfying (above 3.8 out of 5) in all the experimental conditions. Still, their objective self-disclosure (below 1 out of 2) is not as good as the subjective measure. The discrepancy between the two measures might be due to our chatbot's limited social skills. Since our study has aimed to investigate the effects of psychological assessment design on users' self-disclosure in Open-EQs, we did not incorporate the social characteristics into the chatbot design, such as proactivity (e.g., active listening [44]) and emotional intelligence (e.g., empathetic responses [138]), which, however, could encourage honest self-disclosure during the communication [139].

The interaction style indirectly influences subjective and objective self-disclosure through assessment credibility, while questionnaire length does not (Figure 3). Despite the lack of a main effect of questionnaire length, questionnaire length seems to influence the effect of interaction style on self-disclosure (subjective). Although participants thought the design manipulations of psychological assessment did not significantly influence their willingness to disclose themselves (subjective self-disclosure) for Open-EQs, in practice, they showed more self-disclosure in form-based conditions than in conversation-based conditions. This may imply that the form-based interaction is more favorable than the conversation-based interaction regarding users' self-disclosure in their responses to Open-EQs.

We measured the response quality of Open-EQs from multiple dimensions, and the Form*Middle design leads to the highest response quality index (RQI), and the Form*Short design has the highest informativeness. We argue that perceiving higher assessment credibility in the form-based questionnaire motivates participants who feel lonely to talk with the survey chatbot. Furthermore, the response quality of Open-EQs is highly associated with objective self-disclosure, which aligns with the findings of existing work [21, 140].

6. Limitations

Our study has several limitations that need to be mentioned while interpreting our research findings, including the unbalanced gender distribution, narrow scope of mental health, and limited social communication skills of our chatbot.

First, our primary target group is university students who may suffer from loneliness. To reach a broad audience, we have collaborated with the Counseling and Development Center (CDC) of Hong Kong Baptist University (HKBU) to recruit participants within the university. However, we encountered an imbalance in the gender distribution of our participants, primarily because HKBU has a higher ratio of female students. In addition, existing research suggests that loneliness is more commonly experienced by males than females [141]. However, the analysis of gender as a control variable on all dependent variables did not yield significance. Therefore, the gender imbalance should not significantly impact the generalizability of our findings.

Second, we investigated the design of the psychological assessment only for loneliness because the loneliness scale has three validated length versions, which meets our requirement of manipulating questionnaire lengths as short, middle, and long. Strictly speaking, loneliness is not a mental health issue, but it is closely related to various mental health issues such as anxiety, stress, and depression [142, 143]. Lonely people may behave differently from those who suffer from mental health issues regarding self-disclosure intentions. For example, lonely people are more willing to disclose private information than those connected [144], while individuals with depression and anxiety are associated with lessened emotional self-disclosure [145]. Therefore, further study is needed to validate to what extent our findings on the psychological assessment design can be generalized to a survey chatbot for screening other mental health issues.

Third, our current survey design is that Open-EQs were positioned immediately after Closed-EQs. While this sequential arrangement is common in mental health survey design, there are some alternative methods to mix Open-EQs and Closed-EQs. For example, participants could explain their choices of a Closed-EQ through the following Open-EQ. This highlights the need for further research to explore diverse psychological measurement design approaches in survey chatbots.

Fourth, since we have focused on investigating the impacts of the psychological assessment design on user responses to Open-EQs, our survey chatbot provides relatively unified responses according to the length of users' responses. For example, *"I understand."* or *"Thank you. I really appreciate your input."*. However, some participants expected to receive more meaningful and personalized feedback while conversing with the chatbot. For example, *"...the bot response does not reply authentically according to my response." (P56, Con*Middle)* In the future, we plan to incorporate sophisticated social communication skills, such as active listening [44] and bot self-disclosure [46, 45] into a survey chatbot for mental health. Besides, the chatbot powered by large language models (LLMs) [146], e.g., ChatGPT,[13] has demonstrated an impressive ability to understand and generate natural language in conversation. Therefore, we will consider leveraging the LLMs to generate engaging and empathetic responses so as to improve user engagement in the survey chatbot for mental health.

7. Conclusions

We conducted a field study (N=213) that investigated how two prominent design factors of the psychological assessment (i.e., *interaction style* and *questionnaire length*) influence user responses to the open-ended questions (Open-EQs) in a survey chatbot for mental health. The results indicate that the form-based interaction is more favored than the conversation-based interaction for the psychological assessment regarding users' perceived assessment credibility and self-awareness. The increased assessment credibility could further stimulate more self-disclosure and quality responses in Open-EQs. Moreover, although the questionnaire length has a limited impact on user responses to Open-EQs, we suggest that the questionnaire length could be adapted to the assessment purpose and content or be determined based on participants' time pressure. To the best of our knowledge, most existing works on mental health chatbots focus on enhancing chatbots' communication skills to increase user engagement and response quality [44, 21, 46, 45]. However, little work has investigated the potential effect of the psychological assessment design in a survey chatbot for mental health. Finally, we explain our findings through an SEM model containing all design factors, response quality and self-disclosure in Open-EQs, and the users' perceptions of the survey. By investigating two prominent design factors of the psychological assessment in a survey chatbot for mental health, we believe that the findings could be suggestive for researchers and practitioners to better leverage the chatbot technology for improving the quality and user experience of their mental health survey.

Acknowledgements

This work was supported by the Hong Kong Research Grants Council (RGC) GRF project (RGC/HKBU12201620), the Hong Kong Baptist University IG-FNRA

[13]`https://chat.openai.com/chat`

project (RC-FNRA-IG/21-22/SCI/01), and the Hong Kong Baptist University Start-up Grant (RC-STARTUP/21-22/23). In addition, we want to thank our collaborators working in the Counseling and Development Center (CDC) of HKBU, Ms. Vicki Kwan, Ms. Chloe Li, and Ms. Wendy Cheung, and all the participants who participated in our study.

References

[1] D. Eisenberg, Countering the troubling increase in mental health symptoms among us college students, Journal of Adolescent Health 65 (5) (2019) 573–574.

[2] R. Mojtabai, M. Olfson, B. Han, National trends in the prevalence and treatment of depression in adolescents and young adults, Pediatrics 138 (6) (2016).

[3] R. D. Goodwin, A. H. Weinberger, J. H. Kim, M. Wu, S. Galea, Trends in anxiety among adults in the united states, 2008–2018: Rapid increases among young adults, Journal of psychiatric research 130 (2020) 441–446.

[4] T. Gagné, I. Schoon, A. Sacker, Trends in young adults' mental distress and its association with employment: Evidence from the behavioral risk factor surveillance system, 1993–2019, Preventive Medicine 150 (2021) 106691.

[5] N. R. Magson, J. Y. Freeman, R. M. Rapee, C. E. Richardson, E. L. Oar, J. Fardouly, Risk and protective factors for prospective changes in adolescent mental health during the covid-19 pandemic, Journal of youth and adolescence 50 (1) (2021) 44–57.

[6] L. Liang, H. Ren, R. Cao, Y. Hu, Z. Qin, C. Li, S. Mei, The effect of covid-19 on youth mental health, Psychiatric quarterly 91 (3) (2020) 841–852.

[7] D. Courtney, P. Watson, M. Battaglia, B. H. Mulsant, P. Szatmari, Covid-19 impacts on child and youth anxiety and depression: challenges and opportunities, The Canadian Journal of Psychiatry 65 (10) (2020) 688–691.

[8] E. J. Costello, Early detection and prevention of mental health problems: developmental epidemiology and systems of support, Journal of Clinical Child & Adolescent Psychology 45 (6) (2016) 710–717.

[9] J. M. Levitt, N. Saka, L. H. Romanelli, K. Hoagwood, Early identification of mental health problems in schools: The status of instrumentation, Journal of School Psychology 45 (2) (2007) 163–191.

[10] A. A. Abd-Alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, M. Househ, An overview of the features of chatbots in mental health: A scoping review, International Journal of Medical Informatics 132 (2019) 103978.

[11] I. Hungerbuehler, K. Daley, K. Cavanagh, H. G. Claro, M. Kapps, et al., Chatbot-based assessment of employees' mental health: Design process and pilot implementation, JMIR Formative Research 5 (4) (2021) e21678.

[12] A. Schick, J. Feine, S. Morana, A. Maedche, U. Reininghaus, et al., Validity of chatbot use for mental health assessment: Experimental study, JMIR mHealth and uHealth 10 (10) (2022) e28082.

[13] M. E. Te Pas, W. G. Rutten, R. A. Bouwman, M. P. Buise, User experience of a chatbot questionnaire versus a regular computer questionnaire: prospective comparative study, JMIR Medical Informatics 8 (12) (2020) e21982.

[14] S. Kim, J. Lee, G. Gweon, Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–12.

[15] U. Reja, K. L. Manfreda, V. Hlebec, V. Vehovar, Open-ended vs. close-ended questions in web questionnaires, Developments in applied statistics 19 (1) (2003) 159–177.

[16] O. Friborg, J. H. Rosenvinge, A comparison of open-ended and closed questions in the prediction of mental health, Quality & Quantity 47 (3) (2013) 1397–1411.

[17] D. W. Russell, Ucla loneliness scale (version 3): Reliability, validity, and factor structure, Journal of personality assessment 66 (1) (1996) 20–40.

[18] R. Zhou, X. Wang, L. Zhang, H. Guo, Who tends to answer open-ended questions in an e-service survey? the contribution of closed-ended answers, Behaviour & Information Technology 36 (12) (2017) 1274–1284.

[19] I. Borg, C. Zuell, Write-in comments in employee surveys, International Journal of Manpower (2012).

[20] R. M. Poncheri, J. T. Lindberg, L. F. Thompson, E. A. Surface, A comment on employee surveys: Negativity bias in open-ended responses, Organizational Research Methods 11 (3) (2008) 614–630.

[21] Z. Xiao, M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, H. Yang, Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions, ACM Transactions on Computer-Human Interaction (TOCHI) 27 (3) (2020) 1–37.

[22] S. Ganassali, The influence of the design of web survey questionnaires on the quality of responses, in: Survey research methods, Vol. 2, 2008, pp. 21–32.

[23] M. Galesic, M. Bosnjak, Effects of questionnaire length on participation and indicators of response quality in a web survey, Public opinion quarterly 73 (2) (2009) 349–360.

[24] B. Burchell, C. Marsh, The effect of questionnaire length on survey response, Quality and quantity 26 (3) (1992) 233–244.

[25] A. R. Herzog, J. G. Bachman, Effects of questionnaire length on response quality, Public opinion quarterly 45 (4) (1981) 549–559.

[26] L. A. Peplau, D. Perlman, Loneliness: A sourcebook of current theory, research, and therapy, (No Title) (1982).

[27] E. Hards, M. E. Loades, N. Higson-Sweeney, R. Shafran, T. Serafimova, A. Brigden, S. Reynolds, E. Crawley, E. Chatburn, C. Linney, et al., Loneliness and mental health in children and adolescents with pre-existing mental health problems: A rapid systematic review, British Journal of Clinical Psychology 61 (2) (2022) 313–334.

[28] J. Christiansen, P. Qualter, K. Friis, S. Pedersen, R. Lund, C. Andersen, M. Bekker-Jeppesen, M. Lasgaard, Associations of loneliness and social isolation with physical and mental health among adolescents and young adults, Perspectives in public health 141 (4) (2021) 226–236.

[29] J. T. Cacioppo, L. C. Hawkley, R. A. Thisted, Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the chicago health, aging, and social relations study., Psychology and aging 25 (2) (2010) 453.

[30] M. Lasgaard, K. Friis, M. Shevlin, "where are all the lonely people?" a population-based study of high-risk groups across the life span, Social psychiatry and psychiatric epidemiology 51 (2016) 1373–1384.

[31] T. E. Keller, M. Perry, R. Spencer, Reducing social isolation through formal youth mentoring: opportunities and potential pitfalls, Clinical Social Work Journal 48 (2020) 35–45.

[32] N. A. Mayorga, T. Smit, L. Garey, A. K. Gold, M. W. Otto, M. J. Zvolensky, Evaluating the interactive effect of covid-19 worry and loneliness on mental health among young adults, Cognitive therapy and research (2022) 1–9.

[33] K. Cooper, E. Hards, B. Moltrecht, S. Reynolds, A. Shum, E. McElroy, M. Loades, Loneliness, social relationships, and mental health in adolescents during the covid-19 pandemic, Journal of Affective Disorders 289 (2021) 98–104.

[34] S. Marchini, E. Zaurino, J. Bouziotis, N. Brondino, V. Delvenne, M. Delhaye, Study of resilience and loneliness in youth (18–25 years old) during the covid-19 pandemic lockdown measures, Journal of community psychology 49 (2) (2021) 468–480.

[35] A. Martinez, S. Nguyen, The impact of covid-19 on college student well-being (2020).

[36] M. Panayiotou, J. C. Badcock, M. H. Lim, M. J. Banissy, P. Qualter, Measuring loneliness in different age groups: The measurement invariance of the ucla loneliness scale, Assessment 30 (5) (2023) 1688–1715.

[37] K. K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial, JMIR mental health 4 (2) (2017) e7785.

[38] B. Inkster, S. Sarda, V. Subramanian, An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world

data evaluation mixed-methods study, JMIR mHealth and uHealth 6 (11) (2018) e12106.

[39] K. Kroenke, R. L. Spitzer, J. B. Williams, The phq-9: validity of a brief depression severity measure, Journal of general internal medicine 16 (9) (2001) 606–613.

[40] R. L. Spitzer, K. Kroenke, J. B. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: the gad-7, Archives of internal medicine 166 (10) (2006) 1092–1097.

[41] K. Denecke, S. Vaaheesan, A. Arulnathan, A mental health chatbot for regulating emotions (sermo)-concept and usability test, IEEE Transactions on Emerging Topics in Computing 9 (3) (2020) 1170–1182.

[42] R. R. Morris, K. Kouddous, R. Kshirsagar, S. M. Schueller, Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions, Journal of medical Internet research 20 (6) (2018) e10148.

[43] M. Lee, S. Ackermans, N. Van As, H. Chang, E. Lucas, W. IJsselsteijn, Caring for vincent: a chatbot for self-compassion, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13.

[44] Z. Xiao, M. X. Zhou, W. Chen, H. Yang, C. Chi, If i hear you correctly: Building and evaluating interview chatbots with active listening skills, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[45] Y.-C. Lee, N. Yamashita, Y. Huang, W. Fu, " i hear you, i feel you": Encouraging deep self-disclosure through a chatbot, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–12.

[46] Y.-C. Lee, N. Yamashita, Y. Huang, Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional, Proceedings of the ACM on Human-Computer Interaction 4 (CSCW1) (2020) 1–27.

[47] S. Park, A. Thieme, J. Han, S. Lee, W. Rhee, B. Suh, “i wrote as if i were telling a story to someone i knew.”: Designing chatbot interactions for expressive writing in mental health, in: Designing Interactive Systems Conference 2021, 2021, pp. 926–941.

[48] S. Park, J. Choi, S. Lee, C. Oh, C. Kim, S. La, J. Lee, B. Suh, Designing a chatbot for a brief motivational interview on stress management: Qualitative case study, Journal of medical Internet research 21 (4) (2019) e12231.

[49] P. B. Bae Brandtzæg, M. Skjuve, K. K. Kristoffer Dysthe, A. Følstad, When the social becomes non-human: Young people’s perception of social support in chatbots, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–13.

[50] R. G. Maunder, J. J. Hunter, An internet resource for self-assessment of mental health and health behavior: Development and implementation of the self-assessment kiosk, JMIR mental health 5 (2) (2018) e9768.

[51] J. L. Holland, L. M. Christian, The influence of topic interest and interactive probing on responses to open-ended questions in web surveys, Social Science Computer Review 27 (2) (2009) 196–212.

[52] F. G. Conrad, M. P. Couper, R. Tourangeau, M. Galesic, T. Yan, Interactive feedback can improve the quality of responses in web surveys, in: Proc. Surv. Res. Meth. Sect. Am. Statist. Ass, 2005, pp. 3835–3840.

[53] M. Revilla, C. Ochoa, Ideal and maximum length for a web survey., International Journal of Market Research 59 (5) (2017).

[54] P. Edwards, I. Roberts, P. Sandercock, C. Frost, Follow-up by mail in clinical trials: does questionnaire length matter?, Controlled clinical trials 25 (1) (2004) 31–52.

[55] E. Deutskens, K. De Ruyter, M. Wetzels, P. Oosterveld, Response rate and response quality of internet-based surveys: an experimental study, Marketing letters 15 (1) (2004) 21–36.

[56] A. Peytchev, M. P. Couper, S. E. McCabe, S. D. Crawford, Web survey design: Paging versus scrolling, International Journal of Public Opinion Quarterly 70 (4) (2006) 596–607.

[57] U.-D. Reips, Context effects in web surveys, Online social sciences (2002) 69–80.

[58] M. Liu, A. Cernat, Item-by-item versus matrix questions: A web survey experiment, Social Science Computer Review 36 (6) (2018) 690–706.

[59] C. N. Wathen, J. Burkell, Believe it or not: Factors influencing credibility on the web, Journal of the American society for information science and technology 53 (2) (2002) 134–144.

[60] A. P. Chaves, M. A. Gerosa, How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design, International Journal of Human–Computer Interaction 37 (8) (2021) 729–758.

[61] M. Oudejans, L. M. Christian, Using interactive features to motivate and probe responses to open-ended questions, Social and behavioral research and the internet: Advances in applied methods and research strategies (2010) 304–332.

[62] T. Wambsganss, R. Winkler, M. Söllner, J. M. Leimeister, A conversational agent to improve response quality in course evaluations, in: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–9.

[63] S. Rolstad, J. Adler, A. Rydén, Response burden and questionnaire length: is shorter better? a review and meta-analysis, Value in Health 14 (8) (2011) 1101–1108.

[64] D. W. Eby, L. J. Molnar, J. T. Shope, J. M. Vivoda, T. A. Fordyce, Improving older driver knowledge and self-awareness through self-assessment: The driving decisions workbook, Journal of safety research 34 (4) (2003) 371–381.

[65] C. Cook, F. Heath, R. L. Thompson, A meta-analysis of response rates in web-or internet-based surveys, Educational and psychological measurement 60 (6) (2000) 821–836.

[66] Y. Guo, R. W. Proctor, G. Salvendy, What do users want to see? a content preparation study for consumer electronics, in: International Conference on Human-Computer Interaction, Springer, 2009, pp. 413–420.

[67] J. A. Krosnick, Questionnaire design, in: The Palgrave handbook of survey research, Springer, 2018, pp. 439–455.

[68] P. F. Lazarsfeld, The controversy over detailed interviews—an offer for negotiation, Public opinion quarterly 8 (1) (1944) 38–60.

[69] M. Denscombe, The length of responses to open-ended questions: A comparison of online and paper questionnaires in terms of a mode effect, Social Science Computer Review 26 (3) (2008) 359–368.

[70] M. Emde, M. Fuchs, Using adaptive questionnaire design in open-ended questions: A field experiment, in: American Association for Public Opinion Research (AAPOR) 67th Annual Conference, San Diego, USA, 2012, pp. 1–13.

[71] R. H. Dana, T. A. Hoffmann, Health assessment domains: Credibility and legitimization, Clinical Psychology Review 7 (5) (1987) 539–555.

[72] B. Kitchens, C. A. Harle, S. Li, Quality of health-related online search results, Decision Support Systems 57 (2014) 454–462.

[73] Y. Zhang, Searching for specific health-related information in m edline p lus: Behavioral patterns and user experience, Journal of the Association for Information Science and Technology 65 (1) (2014) 53–68.

[74] M. S. Eastin, Credibility assessments of online health information: The effects of source expertise and knowledge of content, Journal of Computer-Mediated Communication 6 (4) (2001) JCMC643.

[75] E. L. Jenkins, J. Ilicic, A. M. Barklamb, T. A. McCaffrey, Assessing the credibility and authenticity of social media content for applications in health communication: scoping review, Journal of medical Internet research 22 (7) (2020) e17296.

[76] X. Zhao, L. Chen, Y. Jin, X. Zhang, Comparing button-based chatbots with webpages for presenting fact-checking results: A case study of health information, Information Processing & Management 60 (2) (2023) 103203.

[77] L. Sbaffi, J. Rowley, Trust and credibility in web-based health information: a review and agenda for future research, Journal of medical Internet research 19 (6) (2017) e218.

[78] A. L. Pieterse, M. Lee, A. Ritmeester, N. M. Collins, Towards a model of self-awareness development for counselling and psychotherapy training, Counselling Psychology Quarterly 26 (2) (2013) 190–207.

[79] K. D. Killian, Development and validation of the emotional self-awareness questionnaire: A measure of emotional intelligence, Journal of Marital and Family Therapy 38 (3) (2012) 502–514.

[80] M. Csikszentmihalyi, M. Csikzentmihaly, Flow: The psychology of optimal experience, Vol. 1990, Harper & Row New York, 1990.

[81] A. Lin, S. Gregor, M. Ewing, Developing a scale to measure the enjoyment of web experiences, Journal of Interactive Marketing 22 (4) (2008) 40–57.

[82] N. Abbas, T. Pickard, E. Atwell, A. Walker, University student surveys using chatbots: Artificial intelligence conversational agents, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 155–169.

[83] C. Ischen, T. Araujo, G. van Noort, H. Voorveld, E. Smit, "i am here to assist you today": The role of entity, interactivity and experiential perceptions in chatbot persuasion, Journal of Broadcasting & Electronic Media 64 (4) (2020) 615–639.

[84] M. C. Han, The impact of anthropomorphism on consumers' purchase decision in chatbot commerce, Journal of Internet Commerce 20 (1) (2021) 46–65.

[85] T. Rietz, I. Benke, A. Maedche, The impact of anthropomorphic and functional chatbot design features in enterprise collaboration systems on user acceptance, in: 14th International Conference on Wirtschaftsinformatik, 2019, pp. 1642–1656.

[86] R. C. Hanna, B. Weinberg, R. P. Dant, P. D. Berger, Do internet-based surveys increase personal self-disclosure?, Journal of Database Marketing & Customer Strategy Management 12 (2005) 342–356.

[87] B. Jacquet, A. Hullin, J. Baratgin, F. Jamet, The impact of the gricean maxims of quality, quantity and manner in chatbots, in: 2019 international conference on information and digital technologies (idt), IEEE, 2019, pp. 180–189.

[88] J. D. Smyth, D. A. Dillman, L. M. Christian, M. McBride, Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality?, Public Opinion Quarterly 73 (2) (2009) 325–337.

[89] R. Flesch, Marks of readable style; a study in adult education., Teachers College Contributions to Education (1943).

[90] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Foundations and Trends® in information retrieval 2 (1–2) (2008) 1–135.

[91] A. Barak, O. Gluck-Ofri, Degree and reciprocity of self-disclosure in online forums, CyberPsychology & Behavior 10 (3) (2007) 407–417.

[92] S. M. Jourard, Self-disclosure, An experimental analysis of the transparent self (1971).

[93] J. H. Kahn, R. M. Hessling, Measuring the tendency to conceal versus disclose psychological distress, Journal of Social and Clinical Psychology 20 (1) (2001) 41–65.

[94] L. C. Miller, J. H. Berg, R. L. Archer, Openers: Individuals who elicit intimate self-disclosure., Journal of personality and social psychology 44 (6) (1983) 1234.

[95] M. Nguyen, Y. S. Bin, A. Campbell, Comparing online and offline self-disclosure: A systematic review, Cyberpsychology, Behavior, and Social Networking 15 (2) (2012) 103–111.

[96] A. N. Joinson, Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity, European journal of social psychology 31 (2) (2001) 177–192.

[97] R. Garett, J. Chiu, L. Zhang, S. D. Young, A literature review: website design and user engagement, Online journal of communication and media technologies 6 (3) (2016) 1.

[98] L. E. Kam, W. G. Chismar, Online self-disclosure: model for the use of internet-based technologies in collecting sensitive health information, International journal of healthcare technology and management 7 (3-4) (2006) 218–232.

[99] M. E. Hughes, L. J. Waite, L. C. Hawkley, J. T. Cacioppo, A short scale for measuring loneliness in large surveys: Results from two population-based studies, Research on aging 26 (6) (2004) 655–672.

[100] P. M. Desmet, M. H. Vastenburg, N. Romero, Mood measurement with pick-a-mood: review of current methods and design of a pictorial self-report scale, Journal of Design Research 14 (3) (2016) 241–279.

[101] Y. Rosseel, The lavaan tutorial, Department of Data Analysis: Ghent University (2014).

[102] e. a. Hair, Joseph F., Multivariate Data Analysis: A Global Perspective. 7th ed., Upper Saddle River: Prentice Hall, 2009.

[103] B. Hilligoss, S. Y. Rieh, Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context, Information Processing & Management 44 (4) (2008) 1467–1484.

[104] A. Sutton, Measuring the effects of self-awareness: Construction of the self-awareness outcomes questionnaire, Europe's journal of psychology 12 (4) (2016) 645.

[105] S. Pieritz, M. Khwaja, A. A. Faisal, A. Matic, Personalised recommendations in mental health apps: The impact of autonomy and data sharing, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–12.

[106] J. A. McCarty, L. J. Shrum, The measurement of personal values in survey research: A test of alternative rating procedures, Public Opinion Quarterly 64 (3) (2000) 271–298.

[107] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.

[108] L. Dybkjær, N. O. Bernsen, H. Dybkjær, Grice incorporated: cooperativity in spoken dialogue, in: Proceedings of the 16th conference on Computational linguistics-Volume 1, 1996, pp. 328–333.

[109] G. N. Leech, 100 million words of english: the british national corpus (bnc), Language Research (1992).

[110] K. Hofland, S. Johansson, Word frequencies in british and american english, Norwegian computing centre for the Humanities, 1982.

[111] E. N. Forsythand, C. H. Martell, Lexical and discourse analysis of online chat dialog, in: International Conference on Semantic Computing (ICSC 2007), IEEE, 2007, pp. 19–26.

[112] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[113] K. Kays, K. Gathercoal, W. Buhrow, Does survey format influence self-disclosure on sensitive question items?, Computers in Human Behavior 28 (1) (2012) 251–256.

[114] R. C. MacCallum, J. T. Austin, Applications of structural equation modeling in psychological research, Annual review of psychology 51 (1) (2000) 201–226.

[115] T. A. Kyriazos, et al., Applied psychometrics: sample size and sample power considerations in factor analysis (efa, cfa) and sem in general, Psychology 9 (08) (2018) 2207.

[116] J. Wang, X. Wang, Sample size for structural equation modeling, Structural equation modeling: Applications using Mplus (2012) 391–428.

[117] J. J. Hoogland, A. Boomsma, Robustness studies in covariance structure modeling: An overview and a meta-analysis, Sociological Methods & Research 26 (3) (1998) 329–367.

[118] K. A. Bollen, R. H. Hoyle, Latent variables in structural equation modeling, Handbook of structural equation modeling (2012) 56–67.

[119] L.-t. Hu, P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural equation modeling: a multidisciplinary journal 6 (1) (1999) 1–55.

[120] P. M. Bentler, D. G. Bonett, Significance tests and goodness of fit in the analysis of covariance structures., Psychological bulletin 88 (3) (1980) 588.

[121] E. C. Norton, B. E. Dowd, M. L. Maciejewski, Marginal effects—quantifying the effect of changes in risk factors in logistic regression models, Jama 321 (13) (2019) 1304–1305.

[122] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative research in psychology 3 (2) (2006) 77.

[123] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, S. A. Murphy, Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support, Annals of Behavioral Medicine 52 (6) (2018) 446–462.

[124] J. P. Chung, W.-s. Yeung, Staff mental health self-assessment during the covid-19 outbreak, East Asian Archives of Psychiatry 30 (1) (2020) 34.

[125] C. R. Colvin, J. Block, D. C. Funder, Overly positive self-evaluations and personality: negative implications for mental health., Journal of personality and social psychology 68 (6) (1995) 1152.

[126] F. Heylighen, J.-M. Dewaele, Formality of language: definition, measurement and behavioral determinants, Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel 4 (1999).

[127] K. M. Rennekamp, P. Witz, Linguistic formality and perceived engagement–investors' reactions to two unique characteristics of social media disclosures, SSRN (2017).

[128] M. A. Hamilton, Message variables that mediate and moderate the effect of equivocal language on source credibility, Journal of Language and Social Psychology 17 (1) (1998) 109–143.

[129] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, G. Mark, Style and alignment in information-seeking conversation, in: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, 2018, pp. 42–51.

[130] G. Eysenbach, J. Powell, O. Kuss, E.-R. Sa, Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review, Jama 287 (20) (2002) 2691–2700.

[131] K. Kroenke, R. L. Spitzer, J. B. Williams, B. Löwe, The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review, General hospital psychiatry 32 (4) (2010) 345–359.

[132] D.-H. Shin, S. Lee, Y. Hwang, How do credibility and utility play in the user experience of health informatics services?, Computers in Human Behavior 67 (2017) 292–302.

[133] R. Lederman, H. Fan, S. Smith, S. Chang, Who can you trust? credibility assessment in online health forums, Health Policy and Technology 3 (1) (2014) 13–25.

[134] H. Gaffney, W. Mansell, S. Tai, et al., Conversational agents in the treatment of mental health problems: mixed-method systematic review, JMIR mental health 6 (10) (2019) e14166.

[135] I. Altman, D. A. Taylor, Social penetration: The development of interpersonal relationships., Holt, Rinehart & Winston, 1973.

[136] L. R. Wheeless, J. Grotz, The measurement of trust and its relationship to self-disclosure, Human Communication Research 3 (3) (1977) 250–257.

[137] P. C. Cozby, Self-disclosure: a literature review., Psychological bulletin 79 (2) (1973) 73.

[138] S. Devaram, Empathic chatbot: Emotional intelligence for empathic chatbot: Emotional intelligence for mental health well-being, arXiv preprint arXiv:2012.09130 (2020).

[139] G. M. Lucas, J. Gratch, A. King, L.-P. Morency, It's only a computer: Virtual humans increase willingness to disclose, Computers in Human Behavior 37 (2014) 94–100.

[140] A. J. Bush, A. Parasuraman, Assessing response quality. a self-disclosure approach to assessing response quality in mall intercept and telephone interviews, Psychology & Marketing 1 (3-4) (1984) 57–71.

[141] M. Barreto, C. Victor, C. Hammond, A. Eccles, M. T. Richins, P. Qualter, Loneliness around the world: Age, gender, and cultural differences in loneliness, Personality and Individual Differences 169 (2021) 110066.

[142] W. D. Killgore, S. A. Cloonan, E. C. Taylor, N. S. Dailey, Loneliness: A signature mental health concern in the era of covid-19, Psychiatry research 290 (2020) 113117.

[143] T. Richardson, P. Elliott, R. Roberts, Relationship between loneliness and mental health in students, Journal of Public Mental Health (2017).

[144] Y. Al-Saggaf, S. Nielsen, Self-disclosure on facebook among female users and its relationship to feelings of loneliness, Computers in Human Behavior 36 (2014) 460–468.

[145] J. H. Kahn, A. M. Garrison, Emotional self-disclosure and emotional avoidance: Relations with symptoms of depression and anxiety., Journal of counseling psychology 56 (4) (2009) 573.

[146] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

Appendix A. Open-ended Questions

Table A.4: The Open-Ended Questions Asked During the Interview Session

ID	Question
Open-EQ1	*In general, how would you describe your current mood?*
Open-EQ2	*What do you think of the influence of the COVID-19 pandemic on your study and life?*
Open-EQ3	*Can you tell me a little bit about any contact you have had with friends or family recently?*
Open-EQ4	*What have you tried to manage isolation and loneliness during COVID-19?*
Open-EQ5	*What do you think could be the main factors contributing to loneliness?*
Open-EQ6	*What would it take for you to feel happier or more at peace?*
Open-EQ7	*Think of something that you feel happy and grateful for, great or small (e.g., the food you eat or the place you live in).*

Appendix B. Post-study Questions

Table B.5: The Questions Asked in the Post-Study

ID	Question
Post-Q1	*What do you think of answering the questions to know your loneliness score?*
Post-Q2	*What do you think of knowing your mental status by chatting with such a bot?*
Post-Q3	*What do you think of answering the questions in conversation with the Percy bot instead of filling in an online form?*
Post-Q4	*What do you think of describing your feelings through talking with the Percy bot?*
Post-Q5	*What questions that the Percy bot asked may make you feel concerned about?*

Appendix C. Descriptive Statistics of Dependent Variables

Table C.6: Descriptive Statistics of Dependent Variables

Dependent variable	**Form*Short** (N=34) Mean (SD)	**Form*Middle** (N=36) Mean (SD)	**Form*Long** (N=35) Mean (SD)	**Con*Short** (N=33) Mean (SD)	**Con*Middle** (N=35) Mean (SD)	**Con*Long** (N=40) Mean (SD)
Subjective Experiences						
Assessment Credibility	3.69 (0.66)	**3.89** (0.56)	3.74 (0.67)	3.34 (0.95)	3.53 (0.84)	3.62 (0.77)
Self-Awareness	3.58 (0.66)	3.75 (0.65)	**3.78** (0.66)	3.50 (0.72)	3.67 (0.82)	3.57 (0.67)
Enjoyment	3.83 (0.68)	**3.91** (0.80)	3.78 (0.71)	3.40 (0.73)	3.84 (0.95)	3.83 (0.69)
Response Quality						
Informativeness	**671.0** (458.3)	625.8 (414.4)	618.6(406.3)	547.5 (329.4)	519.3 (359.0)	644.2 (554.5)
Specificity	1.08 (0.47)	**1.10** (0.40)	1.08 (0.41)	0.97 (0.38)	0.94 (0.40)	1.01 (0.40)
Relevance	1.85 (0.21)	**1.90** (0.17)	1.87 (0.18)	1.86 (0.27)	1.82 (0.21)	1.87 (0.23)
Clarity	1.53 (0.32)	1.58 (0.29)	**1.60** (0.27)	1.54 (0.28)	1.47 (0.30)	1.55 (0.33)
RQI	3.41 (2.19)	**3.56** (1.86)	3.46 (1.88)	3.08 (1.68)	2.80 (1.84)	3.27 (1.87)
Self-Disclosure						
Self-Disclosure (sub.)	4.18 (0.68)	3.82 (0.75)	4.00 (0.66)	3.92 (0.68)	**4.24** (0.69)	4.16 (0.57)
Self-Disclosure (obj.)	**0.96** (0.53)	0.94 (0.47)	0.95 (0.50)	0.88 (0.42)	0.78 (0.51)	0.90 (0.46)
Response Length	**60.9** (44.6)	56.2 (39.5)	55.8 (39.4)	48.5 (31.8)	45.8 (34.6)	57.6 (51.5)
Engagement Duration	267.4 (119.4)	**339.8** (277.7)	278.9 (162.8)	261.9 (104.5)	291.0 (184.1)	333.6 (226.0)

Note: 1. RQI is calculated based on specificity, relevance, and clarity by using Formula (2). 2. The highest value of each dependent variable is marked in bold.