Received: 1 September 2021 | Revised: 14 December 2021 | Accepted: 14 December 2021

DOI: 10.1002/eat.23662

ORIGINAL ARTICLE

International Journal of EATING DISORDERS WILEY

Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial

Ellen E. Fitzsimmons-Craft PhD[1] | William W. Chan PsyD[2,3] | Arielle C. Smith[1] | Marie-Laure Firebaugh LMSW[1] | Lauren A. Fowler PhD[1] | Naira Topooco PhD[3,4] | Bianca DePietro BA[1] | Denise E. Wilfley PhD[1] | C. Barr Taylor MD[2,3] | Nicholas C. Jacobson PhD[5]

[1]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA

[2]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California, USA

[3]Center for m[2]Health, Palo Alto University, Palo Alto, California, USA

[4]Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden

[5]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire, USA

Correspondence
C. Barr Taylor, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA.
Email: btaylor@stanford.edu

Funding information
National Eating Disorders Association Feeding Hope Fund grant; National Heart, Lung, and Blood Institute, Grant/Award Number: T32 HL130357; National Institute of Mental Health, Grant/Award Number: K08 MH120341; Swedish Research Council

Action Editor: Tracey Wade

Abstract

Objective: Prevention of eating disorders (EDs) is of high importance. However, digital programs with human moderation are unlikely to be disseminated widely. The aim of this study was to test whether a chatbot (i.e., computer program simulating human conversation) would significantly reduce ED risk factors (i.e., weight/shape concerns, thin-ideal internalization) in women at high risk for an ED, compared to waitlist control, as well as whether it would significantly reduce overall ED psychopathology, depression, and anxiety and prevent ED onset.

Method: Women who screened as high risk for an ED were randomized ($N = 700$) to (1) chatbot based on the StudentBodies© program; or (2) waitlist control. Participants were followed for 6 months.

Results: For weight/shape concerns, there was a significantly greater reduction in intervention versus control at 3- ($d = -0.20$; $p = .03$) and 6-m-follow-up ($d = -0.19$; $p = .04$). There were no differences in change in thin-ideal internalization. The intervention was associated with significantly greater reductions than control in overall ED psychopathology at 3- ($d = -0.29$; $p = .003$) but not 6-month follow-up. There were no differences in change in depression or anxiety. The odds of remaining nonclinical for EDs were significantly higher in intervention versus control at both 3- (OR = 2.37, 95% CI [1.37, 4.11]) and 6-month follow-ups (OR = 2.13, 95% CI [1.26, 3.59]).

Discussion: Findings provide support for the use of a chatbot-based EDs prevention program in reducing weight/shape concerns through 6-month follow-up, as well as in reducing overall ED psychopathology, at least in the shorter-term. Results also suggest the intervention may reduce ED onset.

Public Significance: We found that a chatbot, or a computer program simulating human conversation, based on an established, cognitive-behavioral therapy-based eating disorders prevention program, was successful in reducing women's concerns about weight and shape through 6-month follow-up and that it may actually reduce eating disorder onset. These findings are important because this intervention, which

Ellen E. Fitzsimmons-Craft and William W. Chan contributed equally to the manuscript.

uses a rather simple text-based approach, can easily be disseminated in order to prevent these deadly illnesses.
Trial registration: OSF Registries; https://osf.io/7zmbv

KEYWORDS
anxiety, chatbot, conversational agent, depression, feeding and eating disorders, female, follow-up studies, humans, psychopathology, risk factors

1 | INTRODUCTION

Eating disorders (EDs) are common, disabling problems (Klump, Bulik, Kaye, Treasure, & Tyson, 2009). Prevention of EDs is of the utmost importance given the wide treatment gap that exists once individuals develop EDs (Fitzsimmons-Craft et al., 2019; Kazdin, Fitzsimmons-Craft, & Wilfley, 2017). Fortunately, risk factors for EDs have been identified (Jacobi, Hayward, de Zwaan, Kraemer, & Agras, 2004; Keel & Forney, 2013), such as weight/shape concerns and thin-ideal internalization (i.e., extent to which an individual "buys into" socially defined ideals of attractiveness). One widely studied targeted prevention program, StudentBodies©, an Internet-based program based on cognitive-behavioral therapy delivered over 8 weeks, significantly reduces weight/shape concerns among women at high risk for the onset of an ED (Taylor et al., 2006), and in the highest risk groups, has been shown to reduce ED onset (Taylor et al., 2016). Some guidance from a real-life, human supporter or moderator improves outcomes (Kass et al., 2014), which is a consistent finding in the literature for digital interventions (Baumeister, Reichler, Munzinger, & Lin, 2014; Richards & Richardson, 2012). However, program moderators for StudentBodies© spent an average of 48.8 min per participant over the intervention (Kass et al., 2014). As such, costs to provide the program and associated human moderation to the large number of people at risk for an ED who might benefit make it unlikely that a human-moderated version can be disseminated widely. One possible solution to reducing delivery costs is to program a chatbot, a computer program that simulates conversation with a human, to mimic aspects of human moderation. Chatbots are widely used in industry and have begun to be used in medical settings (Dingler, Kwasnicka, Wei, Gong, & Oldenburg, 2021), although few studies have examined their effectiveness for mental health issues. Systematic reviews in 2019 and 2020 identified only 12–13 studies on the effects of chatbots on mental health outcomes, with positive effects on psychological distress and some other outcomes (Abd-Alrazaq, Rababeh, Alajlani, Bewick, & Househ, 2020; Gaffney, Mansell, & Tai, 2019). None of the included studies addressed EDs or ED prevention, and notably, the chatbots employed were rather varied in their intended duration and approaches. Chatbots hold promise for both EDs prevention and mental health in general compared with other digital mental health interventions given the interactivity provided by chatbots that mimics therapeutic conversations (Gaffney et al., 2019). Further, research has shown chatbots encourage honest disclosure (Lucas, Gratch, King, & Morency, 2014).

Since publication of the aforementioned chatbot reviews, Beilharz, Sukunesan, Rossell, Kulkarni, and Sharp (2021) published a paper on the development of a chatbot, KIT, designed to support people with concerns about body image and eating, as well as their loved ones. KIT provides psychoeducation, information on how to seek help, and coping skills, including strategies for managing social media, mindfulness, and enjoyable movement. The chatbot also includes information for those seeking to help someone else, including psychoeducation, ED warning signs, and how to seek help (Beilharz et al., 2021).

The aim of this study was to test the hypothesis that a chatbot, based on the cognitive-behavioral therapy-based StudentBodies© program, would significantly reduce key ED risk factors (i.e., weight/shape concerns, thin-ideal internalization) in women at high risk for the onset of an ED compared to waitlist control. Secondary aims were to test the hypotheses that the chatbot vs waitlist control would significantly reduce clinical outcomes (i.e., ED psychopathology, depression, anxiety), as well as prevent ED onset.

2 | METHODS

2.1 | Participants and procedure

Participants were recruited through placing ads on social media, asking social media influencers to post about the study, posting flyers, and through referrals from the National Eating Disorders Association (NEDA) online EDs screen (available at https://www.nationaleatingdisorders.org/screening-tool) and other ongoing EDs research studies. The study was described as one testing whether a chatbot could help reduce risk factors for EDs. Potential participants were directed to an online questionnaire administered through Qualtrics. Online informed consent was obtained from all participants. The survey then determined eligibility, and participants provided baseline data. Participants were informed they would also be asked to complete online assessments at 3- and 6-month follow-ups. Inclusion criteria included being 18–30 years old given that a majority of ED cases onset by this time (Ward, Rodriguez, Wright, Austin, & Long, 2019), identifying as female (given high ED risk in females in particular), and screening as high risk for an ED. Participants were excluded if they did not meet the age/gender criteria, were not at risk for an ED, or screened positive for a clinical/subclinical ED. The Stanford-Washington University ED screen (SWED) was used to determine ED diagnostic or risk status (Graham et al., 2019) and

questions from the Weight Concerns Scale (WCS) were included (Killen et al., 1994; Killen et al., 1996). Responses were used to categorize individuals into one of four categories: (1) possible anorexia nervosa (AN), based on body mass index and elevated weight and shape concerns; (2) possible clinical/subclinical ED other than AN, based on binge eating and/or purging behaviors in the past 3 months; (3) high risk for an ED, based on elevated weight and shape concerns (i.e., group eligible for the trial and based on at least one of the following: (a) total WCS score of 47 or greater; (b) endorsement of weight being more important than most things in life or most important on the WCS; or (c) being very afraid or terrified of gaining three pounds on the WCS; and (4) low risk for an ED based on not screening into one of the above categories. The SWED screening algorithm has been validated and used in past research (Fitzsimmons-Craft et al., 2019; Graham et al., 2019). Participants who screened positive for an ED were provided with referral information. Eligible and interested participants were randomized via Qualtrics (using simple randomization) to the intervention group where they were given immediate access to the chatbot via SMS or Facebook Messenger or the waitlist control where they were informed they could access the chatbot 6 months later and given access as such. The chatbot was described to both groups as a fully automated, conversation-based computer program that would deliver a cognitive-behavioral intervention designed to improve body image. An inactive control condition was chosen for this initial test of the efficacy of the chatbot-based intervention, in order to demonstrate potential benefits of the chatbot versus no intervention (Karlsson & Bergmark, 2015). This comparison is particularly meaningful given general lack of access to preventive interventions for EDs and thus represents an important real-world comparison. Participants were remunerated with $5 each for completion of the baseline and 3-month follow-up and with $10 for completion of 6-month follow-up. All procedures involving human subjects/patients were approved by the Palo Alto University Institutional Review Board.

2.2 | Intervention

StudentBodies© was originally designed as an 8-week traditional web-based program, with users being asked to complete one 30-min web-based session each week (Taylor et al., 2006). This content was reworked by the research team for delivery via a chatbot, while retaining the core intervention principles (Mohr et al., 2015). The program was referred to as Body Positive and was delivered by a chatbot named Tessa, developed by a private mental health chatbot company, X2AI. The program consisted of an introduction, covering information about the program, privacy, crisis protocol, and limitations of the chatbot, and eight sessions delivered as rule- or algorithm-based conversations, which rely on human authoring of conversations, covering the following topics which were covered in the original StudentBodies program: challenging the thin body ideal; media literacy; 4Cs (comparisons, conversations, commercials, and clothing); healthy eating; critical comments; exercise; binge eating; and maintenance. That is, conversations based on these topics were programmed into the chatbot, and the chatbot initiated each conversation in the predetermined order. Participants were encouraged to complete two conversations a week and were told that at that rate the program would take about 1 month to complete. Conversations were designed to take about 10 min each. At the end of each conversation, the chatbot let the user know that it would reach out in 2 days to initiate the next session. After a user completed all eight conversations, the chatbot provided a menu of commands for restarting prior conversations as desired. There was no maximum period of access.

The chatbot that delivered and moderated Body Positive was fully automated. In addition to the Body Positive-specific modules, there were other preexisting modules (e.g., crisis module) and functions (i.e., opting-out of chatbot reminders and recognizing/responding to questions) available from the wider X2AI platform that were triggered based on recognized keywords (e.g., "Unsubscribe" or "?") in users' comments. The crisis module, which provided users with a referral to the crisis hotline in case of an emergency, was triggered based on recognized keywords such as "hurting myself." The chatbot conversational dynamics were meant to mimic natural text-based conversations, and communication was synchronous as the chatbot responded to the user within seconds. Other principles that guided the process of modifying the StudentBodies© content for the chatbot included: (1) keeping the length of each chatbot response short to align with texting culture; (2) having the chatbot send infographics to help reinforce ideas and break up text; (3) offering chatbot responses that were designed to convey warmth and to be appropriate for most users; and (4) using emojis, with the goal of making the program more engaging and aligning with current texting culture. In order to avoid reinforcing possibly problematic statements from users (e.g., "I hate my appearance") while still offering an appropriate response, the chatbot generally did not use nonspecific positive responses (e.g., "Wonderful!") and instead responded with more neutral but warm statements. Figure 1 displays intervention screenshots. As depicted there, the chatbot delivered the conversation line-by-line, at times asking the participant direct or open-ended questions related to the topic at hand.

During the conduct of the trial, research team members monitored the performance of the chatbot by reviewing the transcripts between the chatbot and users at least once a month. A total of over 150,000 responses (105,000 from chatbot and 52,129 from users) were reviewed to identify bugs, erroneous or problematic chatbot responses, and conversations that did not flow well. We note that throughout delivery of the chatbot, the core intervention principles remained the same (Mohr et al., 2015), and changes were focused on removal of bugs and improving conversational flow and quality of the chatbot responses. For more information, see Chan et al. (In press).

2.3 | Measures

2.3.1 | Primary outcomes

The WCS (Killen et al., 1994; Killen et al., 1996) is a five-item self-report questionnaire used to assess weight and shape concerns (range 0–100).

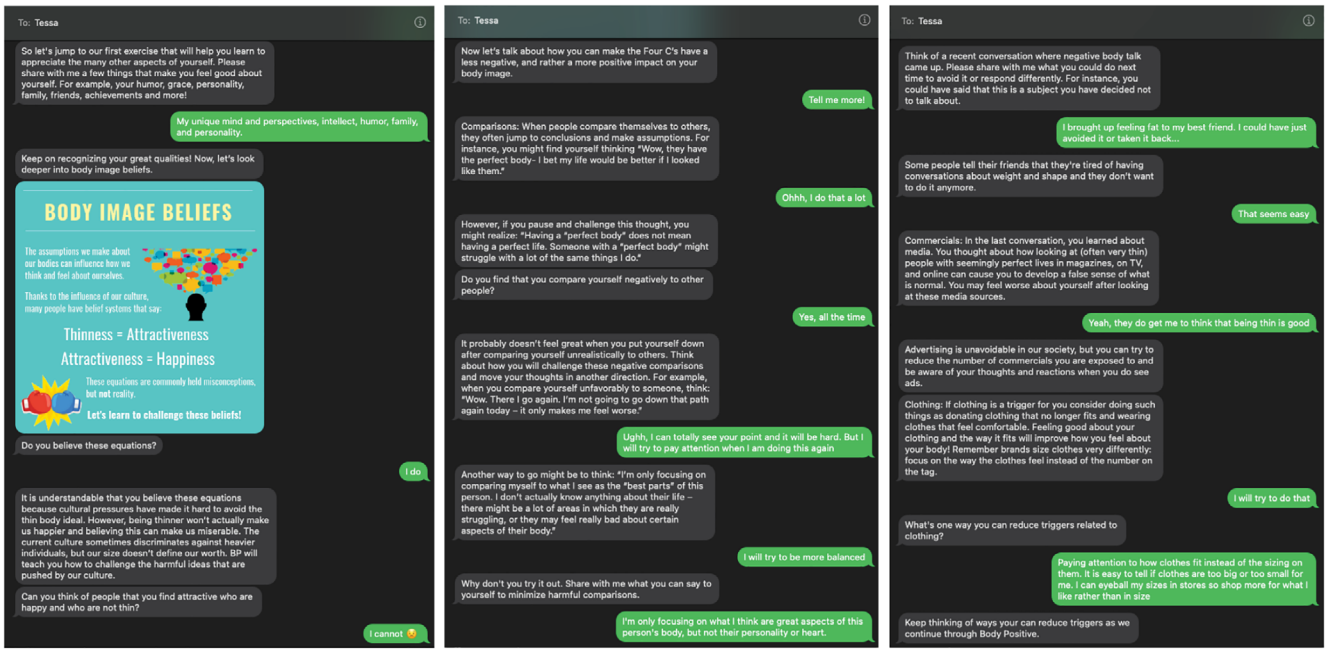

FIGURE 1 Intervention screenshots

The WCS has been shown to be a robust indicator of ED risk (Jacobi, Hayward, et al., 2004). A score of ≥47 has been suggested as a cutoff to indicate increased risk for ED onset (Jacobi, Abascal, & Taylor, 2004). Internal consistency ($\alpha = .79$) was high in this study.

The Internalization: Thin/Low Body Fat subscale of the Sociocultural Attitudes Toward Appearance Questinnaire-4R (SATAQ-4R) is a four-item self-report questionnaire used to assess the cognitive aspect of thin-ideal internalization (Schaefer, Harriger, Heinberg, Soderberg, & Kevin Thompson, 2017). Scores range from 4 to 20. It has good reliability and construct validity in college-age women (Schaefer et al., 2017). Internal consistency ($\alpha = .86$) was high in this study.

2.3.2 | Secondary outcomes

The Eating Disorders Examination-Questionnaire (EDE-Q) is a 28-item self-report questionnaire that is widely used to assess ED attitudes and behaviors (Fairburn & Beglin, 2008). There is a Global score as well as four subscales (Restraint, Eating Concern, Shape Concern, Weight Concern), and all scores range 0–6. A Global score of ≥4 indicates clinical caseness (Luce, Crowther, & Pole, 2008; Quick & Byrd-Bredbenner, 2013). The EDE-Q Global score demonstrated high internal consistency ($\alpha = .91$), as did the subscales (Restraint $\alpha = .79$; Eating Concern $\alpha = .74$; Shape Concern $\alpha = .85$; Weight Concern $\alpha = .68$) in this study.

The Patient Health Questionnaire-8 (PHQ-8) is an eight-item self-report questionnaire that that is widely used to assess current depression, and scores range from 0 to 24 (Kroenke et al., 2009). A PHQ-8 score ≥ 10 has been shown to have 88% sensitivity for identifying major depressive disorder (Kroenke & Spitzer, 2002). Internal consistency ($\alpha = .88$) was high in this study.

The Generalized Anxiety Disorder-7 (GAD-7) is a seven-item self-report questionnaire to measure symptoms of generalized anxiety disorder, and scores range from 0 to 13 (Spitzer, Kroenke, Williams, & Löwe, 2006). A GAD-7 score of ≥10 has been shown to have 89% sensitivity for identifying generalized anxiety disorder (Spitzer et al., 2006). Internal consistency ($\alpha = .92$) was high in this study.

Helpfulness. After each module, participants were asked "Did you find our conversation today helpful?" and were asked to respond using "Yes" or "No."

2.4 | Statistical analysis

Mixed models were used to determine differences in outcomes over time. The model included main effects of group, time (baseline-3-month follow-up and baseline-6-month follow-up), and the interaction between group and time with random intercepts. The sample size was determined partly on previous studies where 75 per group was sufficient to see significant differences with an effect size of $d = .4$, even with only about half of the individuals logging onto the program (Kass et al., 2014). Given the less-intensive nature of the current program, we estimated an effect size of $d = .2$ rather than $d = .4$. Power analyses in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated a sample size of 700 participants ($n = 350$ in each group) would be required to detect between-group differences of $d = .20$. All data were estimated using intent-to-treat analyses, using full information maximum likelihood for missing data (see Figure S1 for a visualization of missing data patterns). Cohen's d was calculated based on $\frac{\beta}{\sqrt{\sigma_0^2+\sigma_1^2}}$. More information on the analytic approach is included in Appendix S1.

2.4.1 | Prevention of ED cases

In addition to analyzing the degree of symptom change continuously, we also evaluated whether the intervention was associated with increased odds of remaining below clinical levels of severity at 3- and 6-month follow-up in the intervention versus control group. For these analyses, we excluded all participants that represented cases at baseline defined based on EDE-Q Global≥4 (Luce et al., 2008; Quick & Byrd-Bredbenner, 2013) and calculated the odds ratio of remaining nonclinical at 3- and 6-month follow-up in the intervention versus control group.

3 | RESULTS

3.1 | Descriptive statistics

Participants were recruited from September 7, 2019 to May 31, 2020, with data collection completed by February 20, 2021. 7,008 individuals were assessed for eligibility, and 700 were eligible and agreed to participate (Figure 2); 352 were randomized to intervention and 348 to control. Follow-up rates can be seen in Figure 2, as well as reasons for withdrawal. Overall completion of at least one follow-up assessment was 62.7% (intervention, 207 of 352 participants [58.8%]; control 232 of 348 participants [66.7%]).

The mean (*SD*) age of the 700 randomized participants was 21.08 (3.09) years. Most identified as White (592 participants [84.6%]), 30 (4.3%) as Asian or South Asian, 23 (3.3%) as Black or African-American, 3 (0.4%) as Native Hawaiian or Other Pacific Islander, 9 (1.3%) as American Indian, 27 (3.9%) as multiracial, and 11 (1.6%) as other races. Regarding ethnicity, 72 (10.3%) identified as Hispanic. In terms of sexual orientation, 377 participants (53.8%) identified as heterosexual, 217 (31.0%) as bisexual, 31 (4.4%) as lesbian, and 15 (2.1%) as asexual. The majority of participants had completed at least some college (502 participants [71.7%]).

There were no significant differences between groups in age (t (693.85) = 0.44, p = .661), race ($\chi 2$ (5) =1.61, p = .901), ethnicity ($\chi 2$ (1) = 2.05, p = .152), education ($\chi 2$ (6) = 1.60, p = .952), or sexual orientation ($\chi 2$ (3) = 3.05, p = .385). Similarly, there were no significant between groups at baseline in the WCS (t(693.62) = −0.68, p = .499), EDE-Q Global (t(695.88) = −1.41, p = .158), EDE-Q Restraint (t(696.98) = −1.15, p = .251), EDE-Q Eating Concern (t(695.81) = −1.40, p = .163), EDE-Q Shape Concern (t(696.99) = −0.88, p = .378),

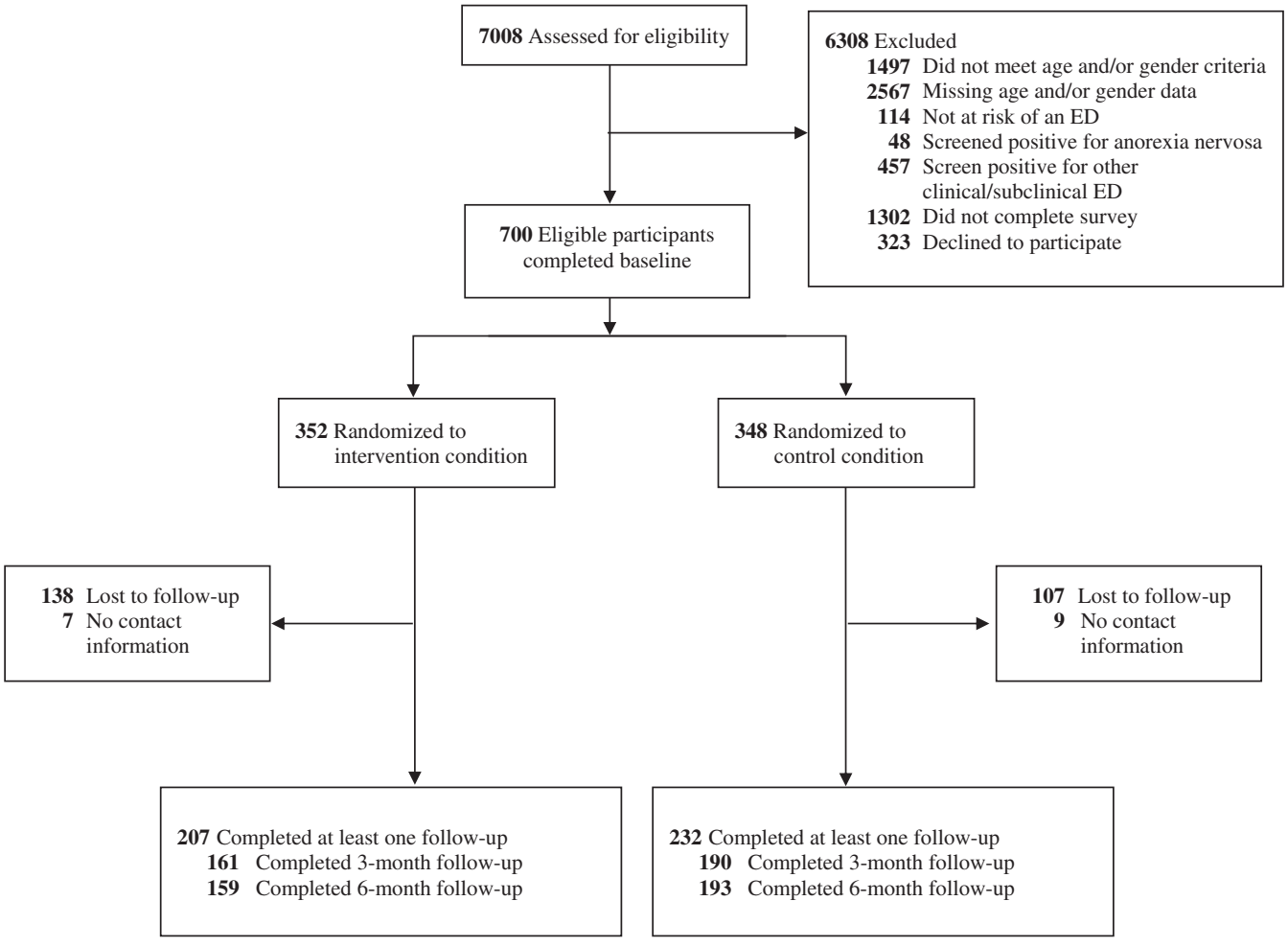

FIGURE 2 Participant flow diagram

EDE-Q Weight Concern ($t(690.67) = -1.12$, $p = .265$), SATAQ-4R subscale ($t(693.68) = -1.76$, $p = .079$), PHQ-8 ($t(696.27) = -0.51$, $p = .609$), or GAD-7 ($t(696.89) = -0.57$, $p = .570$).

3.2 | Outcomes

Table 1 describes descriptive statistics for the outcome variables. Table 2 summarizes model effects.[1] There was a significant interaction between group and degree of change between baseline to 3-month follow-up and baseline to 6-month follow-up for weight/shape concerns (WCS), suggesting that there was significantly more change in the intervention versus control group over both time periods with small effect sizes (see Figure 3). There were no significant differences in degree of change in thin-ideal internalization (SATAQ-4R subscale) between groups over either time period, but there was a significant reduction in both groups from baseline to 6-month follow-up (though not from baseline to 3-month follow-up).

Regarding clinical outcomes, there was a significantly greater reduction in global ED psychopathology (as well as the EDE-Q Eating Concern, Shape Concern, and Weight Concern subscales) in the intervention versus control groups from baseline to 3-month follow-up with small to medium effect sizes, but the differences were no longer significant from baseline to 6-month follow-up (see Figure 3). Nevertheless, both groups showed significant reductions from baseline to 6-month follow-up. There were no significant differences in degree of change in EDE-Q Restraint between groups over either time period, although both groups showed significant reductions at both follow-ups. Similarly, there were no significant differences in degree of change in depression symptoms (PHQ-8) between groups over either time period, but both groups experienced a significant reduction in symptoms over both time periods. Likewise, there were no significant differences in degree of change in anxiety symptoms (GAD-7), but both groups evidenced significant reductions in anxiety symptoms from baseline to 3-month follow-up. Note that there were no significant differences in change between 3- to 6-month follow-up between groups on any variable (see Table S1).

3.3 | Preventing EDs

A total of 114 participants in the control and 104 in the intervention were above clinical levels at baseline, and as such, were excluded from prevention analyses. The odds of remaining nonclinical at 3-month follow-up were significantly higher in the intervention group compared to the control group (OR = 2.37, 95% CI [1.37, 4.11]). Specifically, in the observed data, there was an ED incidence rate of 18.8% in the control and 8.9% in the intervention at 3-month follow-up. Similarly, the odds of remaining nonclinical through the 6-month follow-up were significantly higher in the intervention compared to the control (OR = 2.13, 95% CI [1.26, 3.59]), with the observed data suggesting that there was an ED incidence rate of 19.3% in the control and 10.5% in the intervention at 6-month follow-up.

3.4 | Engagement

Of the 352 participants assigned to the intervention condition, 210 (59.7%) inputted an ID into Tessa (allowing them to use the platform). These 210 users interacted with Tessa for an average of four total interactions (median = 2 interactions) and spent an average of 11 min per interaction (median = 8 min). These interactions occurred across an average of 16 days (median = 6 days).

3.5 | Helpfulness

The percentage of participants who rated each conversation as helpful was 66.2% for conversation 1, 72.3% for conversation 2, 73.8% for

TABLE 1 Outcomes for participants in the intervention condition compared with the control condition

	Baseline		3-month follow-up		6-month follow-up	
Measure	Intervention *M* (*SD*)	Control *M* (*SD*)	Intervention *M* (*SD*)	Control *M* (*SD*)	Intervention *M* (*SD*)	Control *M* (*SD*)
Weight concerns scale	70.59 (15.12)	71.34 (13.90)	61.12 (19.08)	65.57 (17.27)	60.80 (20.55)	63.99 (17.30)
SATAQ-4R thin/low body fat subscale	16.36 (3.56)	16.82 (3.27)	15.48 (3.62)	16.76 (3.07)	15.35 (3.94)	16.11 (3.84)
EDE-Q global	3.38 (1.04)	3.50 (1.07)	2.73 (1.31)	3.26 (1.25)	2.77 (1.34)	3.03 (1.24)
EDE-Q restraint	2.67 (1.58)	2.81 (1.55)	2.33 (1.69)	2.67 (1.69)	2.34 (1.72)	2.37 (1.69)
EDE-Q eating concern	2.32 (1.34)	2.46 (1.38)	1.74 (1.35)	2.37 (1.45)	1.79 (1.43)	2.13 (1.39)
EDE-Q shape concern	4.49 (1.14)	4.57 (1.13)	3.63 (1.49)	4.19 (1.31)	3.63 (1.53)	3.97 (1.37)
EDE-Q weight concern	4.06 (1.08)	4.15 (1.17)	3.25 (1.48)	3.81 (1.34)	3.33 (1.48)	3.68 (1.39)
PHQ-8	13.74 (5.93)	13.97 (5.66)	11.30 (6.14)	11.92 (5.68)	11.09 (6.42)	11.67 (5.55)
GAD-7	11.94 (5.95)	12.20 (5.94)	10.12 (6.12)	10.77 (5.71)	10.40 (6.14)	10.96 (5.51)

Abbreviations: EDE-Q, Eating Disorder Examination-Questionnaire; GAD-7, Generalized Anxiety Disorder-7; PHQ-8, Patient Health Questionnaire-8; SATAQ-4R, Sociocultural Attitudes Toward Appearance Questinnaire-4R; WCS, Weight Concerns Scale.

TABLE 2 Estimated effects of the intervention on outcome measures

Outcome	Main effect of Baseline-3-month follow-up time	Main effect of Baseline-6-month follow-up time	Interaction between Baseline-3-month follow-up time and group assignment	Interaction between Baseline-6-month follow-up time and group assignment
	β_2 (*SE*), *d*, *p*	β_4 (*SE*), *d*, *p*	β_3 (*SE*), *d*, *p*	β_5 (*SE*), *d*, *p*
Weight concerns scale	−6.30 (1.06), $d = -0.38$, $p < .001^*$	−7.43 (1.07), $d = -0.45$, $p < .001^*$	−3.37 (1.56), $d = -0.20$, $p = .031^*$	−3.15 (1.56), $d = -0.19$, $p = .044^*$
SATAQ-4R thin/low body fat subscale	−0.15 (0.23), $d = -0.04$, $p = .50$	−0.73 (0.23), $d = -0.21$, $p = .001^*$	−0.62 (0.33), $d = -0.18$, $p = .06$	−0.22 (0.33), $d = -0.06$, $p = .51$
EDE-Q global	−0.26 (0.07), $d = -0.22$, $p = .001^*$	−0.44 (0.08), $d = -0.38$, $p < .001^*$	−0.33 (0.11), $d = -0.29$, $p = .003^*$	−0.12 (0.11), $d = -0.10$, $p = .27$
EDE-Q restraint	−0.21 (0.11), $d = -0.13$, $p = .047^*$	−0.39 (0.11), $d = -0.24$, $p < .001^*$	−0.18 (0.16), $d = -0.11$, $p = .252$	0.04 (0.16), $d = 0.02$, $p = .82$
EDE-Q eating concern	−0.13 (0.09), $d = -0.10$, $p = .13$	−0.36 (0.09), $d = -0.26$, $p < .001^*$	−0.39 (0.13), $d = -0.28$, $p = .003^*$	−0.10 (0.13), $d = -0.08$, $p = .43$
EDE-Q shape concern	−0.36 (0.08), $d = -0.28$, $p < .001^*$	−0.56 (0.08), $d = -0.45$, $p < .001^*$	−0.38 (0.12), $d = -0.3$, $p = .002^*$	−0.24 (0.12), $d = -0.19$, $p = .05$
EDE-Q weight concern	−0.32 (0.09), $d = -0.26$, $p < .001^*$	−0.44 (0.09), $d = -0.35$, $p < .001^*$	−0.36 (0.13), $d = -0.28$, $p = .005^*$	−0.19 (0.13), $d = -0.15$, $p = .14$
PHQ-8	−1.21 (0.35), $d = -0.20$, $p = .001^*$	−1.53 (0.36), $d = -0.26$, $p < .001^*$	−0.55 (0.52), $d = -0.09$, $p = .29$	−0.70 (0.52), $d = -0.12$, $p = .18$
GAD-7	−0.77 (0.38), $d = -0.13$, $p = .044^*$	−0.65 (0.39), $d = -0.11$, $p = .09$	−0.05 (0.56), $d = -0.01$, $p = .93$	−0.32 (0.57), $d = -0.05$, $p = .57$

Abbreviations: EDE-Q, Eating Disorder Examination-Questionnaire; GAD-7, Generalized Anxiety Disorder-7; PHQ-8, Patient Health Questionnaire-8; SATAQ-4R, Sociocultural Attitudes Toward Appearance Questinnaire-4R; WCS, Weight Concerns Scale.

$^*p < .05$.

conversation 3, 79.4% for conversation 4, 65.0% for conversation 5, 68.8% for conversation 6, 80.5% for conversation 7, and 83.3% for conversation 8.

4 | DISCUSSION

Our chatbot-based EDs prevention program was associated with significantly greater reductions in weight/shape concerns versus waitlist control at both 3- and 6-month follow-ups. The controlled effect sizes ($ds = -.20$ and $-.19$), are in line with the meta-analytic findings on the effects of other digital ED prevention programs on weight/shape concerns (Linardon, Shatte, Messer, Firth, & Fuller-Tyszkiewicz, 2020). While there was a significant reduction across groups on thin-ideal internalization from baseline to 6-month follow-up, there were no significant differences in the magnitude of these changes between groups across either time period, suggesting future iterations may need to include additional content to target this construct, for example, additional exercises to encourage consideration of negative effects of pursuing the thin-ideal in order to induce cognitive dissonance (Becker & Stice, 2017; Stice & Presnell, 2007).

Regarding secondary outcomes, results suggested that the intervention was associated with a significantly greater reduction in global ED psychopathology compared to the control group from baseline to 3-month follow-up, with a small to medium effect size ($d = -.29$), again, in line with the meta-analytic findings on effects of other digital ED prevention programs on these concerns (Linardon et al., 2020). However, these differences did not remain significant at 6-month follow-up. When examining the subscales of the EDE-Q, there was likewise a significantly greater reduction in eating concern, shape concern, and weight concern in the intervention versus control groups at 3-month follow-up, with small-to-medium-effect sizes; however, there were no differences between groups in change in restraint. Although both groups demonstrated significant reductions in depression and anxiety (although only from baseline to 3-month follow-up in the case of anxiety), there were no differences in magnitude of change across groups. These concerns started out and remained at clinical levels over follow-up.

The odds of remaining nonclinical for EDs were significantly higher in the intervention group compared to the control group at both 3 and 6 months. The rate of ED onset in the control group by 6 months was 19.3% compared to 10.5% in the intervention group. However, these results need to be interpreted with caution. First, incidence rates are quite high compared to other studies. For instance, in a previous study of college-age women at risk for EDs who were also identified based on presence of elevated weight/shape concerns (Taylor et al., 2006), we found an overall ED onset rate of just 6.6% in the control at 1 year. Second, we defined caseness in this study based on a self-report measure which is likely to inflate rates, compared to the use of clinical interview in other studies, including the one referenced above (Taylor et al., 2006). Notably, the baseline EDE-Q Global mean scores, which were used for censoring cases, were 3.38 (1.04) in the intervention and 3.50 (1.07) in the control compared to

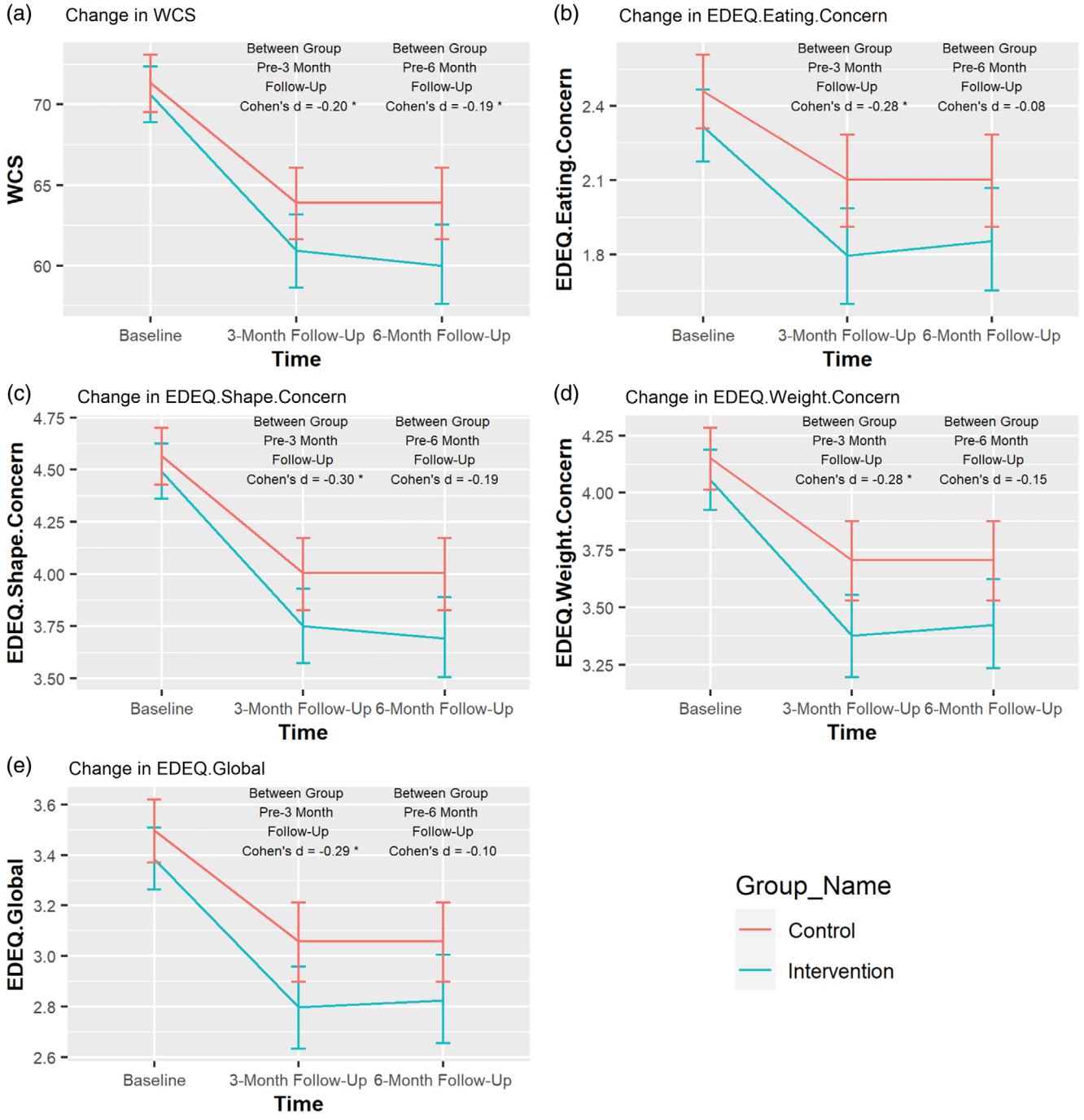

FIGURE 3 Trajectories of the weight concerns scale and eating disorder examination-questionnaire scores. *Note*. Error bars denote 95% CIs. (a) Change in WCS; (b) Change in EDEQ.Eating.Concern; (c) Change in EDEQ.Shape.Concern; (d) Change in EDEQ.Weight.Concern; (e) Change in EDEQ.Global

about 2.6 in our previous study (Taylor et al., 2006). Third, the WCS scores in this study remained high even after intervention, even though there was a significant change and greater reductions in the intervention versus control group. For instance, the 6-month mean WCS in this study in the intervention group (60.80 [20.55]) was the same as the baseline WCS in the control group (60.5 [13.5]) in our previous study (Taylor et al., 2006). All these data suggest we recruited a very high-risk sample. Nevertheless, if confirmed in other studies, the data suggest that a preventive chatbot may reduce ED onset.

Helpfulness ratings varied across the conversations. It is possible that latter conversations (e.g., conversations 7 and 8) were rated as the most helpful as these conversations were only completed by (and thus helpfulness ratings were only provided by) users that made it this far into the program. It is also notable that conversation 4, which addresses regular eating and challenges myths about diets—key

components of CBT for EDs (Fairburn, 2008), was rated as particularly helpful. Future research may wish to further explore user perceptions of conversation helpfulness to inform program refinements.

Strengths of this study include the large number of participants recruited across the United States and broad inclusion criteria. A key limitation is that engagement could be improved, with only 60% of intervention participants starting the chatbot. Once they started engaging, participants interacted with the chatbot four times on average (~11 min each time), over an average of 16 days. Engagement is a known challenge in digital mental health, including ED prevention programs (Linardon et al., 2020) and chatbots (Gaffney et al., 2019). Future work should address the issue of improving engagement with mental health chatbots, including for ED prevention, which may include testing the impact of increasing the interactive nature of the chatbot through artificial intelligence on engagement and outcomes, which is being increasingly used by commercially available chatbots, such as Woebot, which does not currently address EDs (Prochaska et al., 2021). It is also possible that initial uptake of the chatbot could be improved in the future through even more streamlined enrollment processes. Notably, even with the level of engagement observed in the current study, the intervention group demonstrated significant improvement versus control on key metrics. Another limitation is lack of racial/ethnic diversity in the sample. However, there was diversity with regard to sexual orientation. We also note the high number of individuals excluded for not meeting our age/gender criteria, suggesting potential interest from individuals from other demographic groups in similar ED prevention approaches in the future. For example, 17% of those who were assessed for eligibility were 11–17 years old. Further, it is notable that the mean age of the sample in this study was low (21 years) despite targeting women up to the age of 30, suggesting particular interest from young women. Finally, there was a high study attrition rate. However, our attrition was comparable to that observed for long-term follow-up (>8 weeks; 36%) in studies of smartphone-delivered interventions for mental health problems identified in a meta-analytic review (Linardon & Fuller-Tyszkiewicz, 2020), and our analytic approach was able to capitalize on available data. There are other important future directions as well. For example, future research might wish to compare the chatbot to more active control conditions (e.g., chatbot providing generic mental health support, online StudentBodies© intervention with human moderation). Another key future direction may involve assessment of the quality and/or length of participant responses and how this may relate to outcome. Finally, there may be value in further comparing the approaches used by Tessa and the other existing body image-related chatbot, KIT, and systematically testing which elements may be most successful in improving body image and reducing ED concerns.

Overall, findings provide support for the use of a fully automated, highly disseminable chatbot-based EDs prevention program in reducing weight/shape concerns, one of the most robust risk factors for onset for an ED (Jacobi, Hayward, et al., 2004; Keel & Forney, 2013), through 6-month follow-up, as well as in reducing overall ED psychopathology, at least in the shorter term. Results also suggested the intervention may reduce ED onset. Future research could work to further improve engagement and efficacy, including addressing thin-ideal internalization and reducing comorbid concerns, with the ultimate goal being to robustly prevent ED onset. Excitingly, this intervention has high potential for implementation in the "real world," such as through ongoing deployment through our technology partner, a private mental health chatbot company, X2AI. As one option for reaching those in need with this intervention, the chatbot could be made available through NEDA, including through their online EDs screen. The NEDA online screen is accessed by over 200,000 respondents per year, the majority of whom screen positive or at high risk for an ED (Fitzsimmons-Craft, Balantekin, Graham, et al., 2019). Given the high disseminability of the intervention, based on its rather simple text-based approach, there may be opportunities for additional dissemination through other nonprofit organizations or social media outlets as well. Future research should evaluate results of various real-world implementation efforts.

ACKNOWLEDGMENTS

This study was supported by the NEDA Feeding Hope Fund. This study was also supported by National Institute of Mental Health (K08 MH120341), the National Heart, Lung, and Blood Institute (T32 HL130357), and the Swedish Research Council. We sincerely thank our technology partner, X2AI, for their support, without whom this work would not have been possible.

CONFLICTS OF INTEREST

The authors do not have conflicts to declare.

DATA AVAILABILITY STATEMENT

Data are available upon reasonable request to the corresponding author.

ORCID

Ellen E. Fitzsimmons-Craft https://orcid.org/0000-0001-7064-3835
C. Barr Taylor https://orcid.org/0000-0002-4564-6548

ENDNOTE

[1] Note that multicollinearity was checked using variance inflation factors (VIF) of all coefficients. All VIFs were well below 5, suggesting that collinearity is not an issue (Hair, Ringle, & Sarstedt, 2011). Note that the normality of the residuals was checked visually using QQ-plots, and by calculating the skewness and kurtosis. All visual inspections suggested approximate normality of residuals. Additionally, all residuals showed abs(skewness) < 2 and abs(kurtosis) < 7, also suggesting normality of residuals (Byrne, 2013; Hair Jr, Black, Babin, & Anderson, 2010).

REFERENCES

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *22*(7), e16021.

Baumeister, H., Reichler, L., Munzinger, M., & Lin, J. (2014). The impact of guidance on internet-based mental health interventions—A systematic review. *Internet Interventions*, *1*(4), 205–215.

Becker, C. B., & Stice, E. (2017). From efficacy to effectiveness to broad implementation: Evolution of the body project. *Journal of Consulting and Clinical Psychology*, *85*(8), 767–782.

Beilharz, F., Sukunesan, S., Rossell, S. L., Kulkarni, J., & Sharp, G. (2021). Development of a positive body image chatbot (KIT) with young people and parents/Carers: Qualitative focus group study. *Journal of Medical Internet Research*, *23*(6), e27807.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Chan, W. W., Fitzsimmons-Craft, E. E., Smith, A. C., Firebaugh, M., Fowler, L. A., DePietro, B., … Jacobson, N. C. (In Press). Challenges in designing a mental health prevention chatbot: Lessons learned from the field. *JMIR Formative Research*.

Dingler, T., Kwasnicka, D., Wei, J., Gong, E., & Oldenburg, B. (2021). The use and promise of conversational agents in digital health. *Yearbook of Medical Informatics*, *30*, 191–199.

Fairburn, C. G. (2008). *Cognitive behavior therapy and eating disorders*. New York: Guilford Press.

Fairburn, C. G., & Beglin, S. J. (2008). Eating disorder examination questionnaire. In Cognitive Behavior Therapy and Eating Disorders (pp. 309–313). New York: Guilford Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fitzsimmons-Craft, E. E., Balantekin, K. N., Eichen, D. M., Graham, A. K., Monterubio, G. E., Sadeh-Sharvit, S., … Karam, A. M. (2019). Screening and offering online programs for eating disorders: Reach, pathology, and differences across eating disorder status groups at 28 US universities. *International Journal of Eating Disorders*, *52*(10), 1125–1136.

Fitzsimmons-Craft, E. E., Balantekin, K. N., Graham, A. K., Smolar, L., Park, D., Mysko, C., … Wilfley, D. E. (2019). Results of disseminating an online screen for eating disorders across the US: Reach, respondent characteristics, and unmet treatment need. *International Journal of Eating Disorders*, *52*(6), 721–729.

Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Mental Health*, *6*(10), e14166.

Graham, A. K., Trockel, M., Weisman, H., Fitzsimmons-Craft, E. E., Balantekin, K. N., Wilfley, D. E., & Taylor, C. B. (2019). A screening tool for detecting eating disorder risk and diagnostic symptoms among college-age women. *Journal of American College Health*, *67*(4), 357–366.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). New York: Pearson.

Jacobi, C., Abascal, L., & Taylor, C. B. (2004). Screening for eating disorders and high-risk behavior: Caution. *International Journal of Eating Disorders*, *36*(3), 280–295.

Jacobi, C., Hayward, C., de Zwaan, M., Kraemer, H. C., & Agras, W. S. (2004). Coming to terms with risk factors for eating disorders: Application of risk terminology and suggestions for a general taxonomy. *Psychological Bulletin*, *130*(1), 19–65.

Karlsson, P., & Bergmark, A. (2015). Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders. *Addiction*, *110*(3), 420–428.

Kass, A. E., Trockel, M., Safer, D. L., Sinton, M. M., Cunning, D., Rizk, M. T., … Jacobi, C. (2014). Internet-based preventive intervention for reducing eating disorder risk: A randomized controlled trial comparing guided with unguided self-help. *Behaviour Research and Therapy*, *63*, 90–98.

Kazdin, A. E., Fitzsimmons-Craft, E. E., & Wilfley, D. E. (2017). Addressing critical gaps in the treatment of eating disorders. *International Journal of Eating Disorders*, *50*(3), 170–189.

Keel, P. K., & Forney, K. J. (2013). Psychosocial risk factors for eating disorders. *International Journal of Eating Disorders*, *46*(5), 433–439.

Killen, J. D., Taylor, C. B., Hayward, C., Haydel, K. F., Wilson, D. M., Hammer, L., … Strachowski, D. (1996). Weight concerns influence the development of eating disorders: A 4-year prospective study. *Journal of Consulting and Clinical Psychology*, *64*(5), 936–940.

Killen, J. D., Taylor, C. B., Hayward, C., Wilson, D. M., Haydel, K. F., Hammer, L. D., … Varady, A. (1994). Pursuit of thinness and onset of eating disorder symptoms in a community sample of adolescent girls: A three-year prospective analysis. *International Journal of Eating Disorders*, *16*(3), 227–238.

Klump, K. L., Bulik, C. M., Kaye, W. H., Treasure, J., & Tyson, E. (2009). Academy for eating disorders position paper: Eating disorders are serious mental illnesses. *International Journal of Eating Disorders*, *42*(2), 97–103.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. In: SLACK Incorporated, Thorofare, NJ.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Linardon, J., & Fuller-Tyszkiewicz, M. (2020). Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *Journal of Consulting and Clinical Psychology*, *88*(1), 1–13.

Linardon, J., Shatte, A., Messer, M., Firth, J., & Fuller-Tyszkiewicz, M. (2020). E-mental health interventions for the treatment and prevention of eating disorders: An updated systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *88*(11), 994–1007.

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

Luce, K. H., Crowther, J. H., & Pole, M. (2008). Eating disorder examination questionnaire (EDE-Q): Norms for undergraduate women. *International Journal of Eating Disorders*, *41*(3), 273–276.

Mohr, D. C., Schueller, S. M., Riley, W. T., Brown, C. H., Cuijpers, P., Duan, N., … Cheung, K. (2015). Trials of intervention principles: Evaluation methods for evolving behavioral intervention technologies. *Journal of Medical Internet Research*, *17*(7), e166.

Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., & Robinson, A. (2021). A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. *Journal of Medical Internet Research*, *23*(3), e24850.

Quick, V. M., & Byrd-Bredbenner, C. (2013). Eating disorders examination questionnaire (EDE-Q): Norms for US college students. *Eating and Weight Disorders*, *18*(1), 29–35.

Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clinical Psychology Review*, *32*(4), 329–342.

Schaefer, L. M., Harriger, J. A., Heinberg, L. J., Soderberg, T., & Kevin Thompson, J. (2017). Development and validation of the sociocultural attitudes towards appearance questionnaire-4-revised (SATAQ-4R). *International Journal of Eating Disorders*, *50*(2), 104–117.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Stice, E., & Presnell, K. (2007). *The body project: Promoting body acceptance and preventing eating disorders*. New York: Oxford University Press.

Taylor, C. B., Bryson, S., Luce, K. H., Cunning, D., Doyle, A. C., Abascal, L. B., … Wilfley, D. E. (2006). Prevention of eating disorders in at-risk college-age women. *Archives of General Psychiatry*, *63*(8), 881–888.

Taylor, C. B., Kass, A. E., Trockel, M., Cunning, D., Weisman, H., Bailey, J., … Jacobi, C. (2016). Reducing eating disorder onset in a very high risk sample with significant comorbid depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *84*(5), 402–414.

Ward, Z. J., Rodriguez, P., Wright, D. R., Austin, S. B., & Long, M. W. (2019). Estimation of eating disorders prevalence by age and associations with mortality in a simulated nationally representative US cohort. *JAMA Network Open*, *2*(10), e1912925.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Fitzsimmons-Craft, E. E., Chan, W. W., Smith, A. C., Firebaugh, M.-L., Fowler, L. A., Topooco, N., DePietro, B., Wilfley, D. E., Taylor, C. B., & Jacobson, N. C. (2022). Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. *International Journal of Eating Disorders*, *55*(3), 343–353. https://doi.org/10.1002/eat.23662