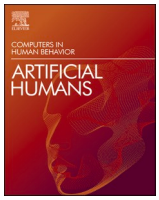
Computers in Human Behavior: Artificial Humans 2 (2024) 100035

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Can robots do therapy?: Examining the efficacy of a CBT bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms

Laura Eltahawy [a], Todd Essig [b], Nils Myszkowski [a], Leora Trub [a,*]

[a] Pace University, United States
[b] William Alanson White Institute of Psychiatry Psychoanalysis and Psychology, United States

ARTICLE INFO

Keywords:
Automated intelligence (AI)
Behavioral intervention technologies (BITs)
Mental health
Telepsychology
Therapy bot
Woebot

ABSTRACT

Artificial intelligence therapy bots are gaining traction in the psychotherapy marketplace. Yet, the only existing study examining the efficacy of a therapy bot lacks any meaningful controls for comparison in claiming its effectiveness to treat depression. The current study aims to examine the efficacy of Woebot against three control conditions, including ELIZA, a basic (non-"smart") conversational bot, a journaling app, and a passive psychoeducation control group. In a sample of 65 young adults, a repeated measures ANOVA failed to detect differences in symptom reduction between active and passive groups. In follow-up analyses using paired samples t-tests, ELIZA users experienced mental health improvements with the largest effect sizes across all mental health outcomes, followed by daily journaling, then Woebot, and finally psychoeducation. Findings reveal that Woebot does not offer benefit above and beyond other self-help behavioral intervention technologies. They underscore that using a no-treatment control group study design to market clinical services should no longer be acceptable nor serve as an acceptable precursor to marketing a chatbot as functionally equivalent to psychotherapy. Doing so creates unnecessary risk for consumers of psychotherapy and undermines the clinical value of robotic therapeutics that could prove effective at addressing mental health problems through rigorous research.

1. The telehealth landscape: benefits and limitations

The pandemic prompted a steep rise in depression, anxiety, and stress, which accelerated a decades long interest in telepsychology, including automated chatbots (Zeavin, 2022). By August 2021, US adult were four times more likely to exhibit symptoms anxiety and depression than they were in 2019, which brought the prevalence of anxiety to 37.2% and depression to 31.1% (DeAngelis, 2021). Alongside orders to stay at home, this influx of mental health needs prompted increased demand for and utilization of remote options for mental health treatment (Garfan et al., 2021). Even before the pandemic, there had been an expansion of therapeutic modalities beyond traditional office-based settings to better accommodate the diverse needs and lifestyles of individuals (Lattie, Stiles-Shields, & Graham, 2022; Zilberstein, 2015). As the pandemic developed and prompted mental health care to transition to an almost fully remote enterprise (Garfan et al., 2021), the benefits of telehealth and their relevance to mental health care became much more apparent (Cao, Zhang, & Liu, 2022; Lattie et al., 2022).

Telepsychology is an umbrella term that refers to the use of technological modalities such as telephone, e-mail, text, videoconferencing, mobile applications, and Web-based programs to deliver psychological services (Drum & Littleton, 2014). For decades, telepsychology has been recognized for widening the provision of services to people with limited resources (Caspar, 2004; Lattie et al., 2022), and those who have personal or circumstantial restrictions in accessing face-to-face psychotherapy due to disability, remote geographical locations or spontaneous events that interfere with travel for therapeutic services (Anthony, 2003). Prior to the pandemic, telepsychology was commonly used to augment in-person services, with examples including the use of teleconferencing for remote therapy sessions, phone apps to track and record moods (Zilberstein, 2015), and computer programs that allowed clinicians to offer diagnostic feedback, cognitive rehabilitation, and other traditionally in-person services (Anthony, 2003). Some telepsychology modes moved away from human contact all together, such as online psychoeducation materials (Gratzer, Strudwick, & Yeung, 2019) and self-administered online therapeutic treatment programs

* Corresponding author.
E-mail address: ltrub@pace.edu (L. Trub).

https://doi.org/10.1016/j.chbah.2023.100035
Received 2 December 2023; Accepted 3 December 2023
Available online 8 December 2023

(Cao et al., 2022; Castelnuovo, Gaggioli, Mantovani, & Riva, 2003; Lattie et al., 2022). These self-help programs, also known as behavioral intervention technologies (BITs), allow people to access online resources aimed at managing their own psychological needs (Lattie et al., 2022; Simmons, Garcia, Howell, & Leong, 2016). They span a wide array of mental health concerns, including phobias, eating disorders, and addiction (Castelnuovo et al., 2003). These programs are often administered through mobile applications, with common examples including mood trackers/monitors and apps that offer access to cognitive behavioral techniques or mindfulness coaching for specific conditions such as anxiety, depression, and Post-Traumatic Stress Disorder (Diano, Sica, & Ponticorvo, 2023; Simmons et al., 2016). BITs have been shown to support behavioral and mental health treatment outcomes by enabling individuals to monitor compliance, track mood, and access psychoeducation materials and relaxation techniques (Simmons et al., 2016). But, as helpful as they were, they were not seen as equivalent to professionally provided psychotherapy; rather, they were often used in an adjunctive fashion (Clough & Casey, 2015). In a recent systematic review of the literature, (Conley et al., 2022) note that this literature has yet to answer the question of whether BITs can be considered an effective and low-cost substitution to traditional psychotherapeutic services, or if they are best used as a way to engage people initially and ultimately connect them to traditional services.

These programs and services have increased the accessibility and scope of mental health resources. But there are concerns about the concurrent decline in the quality of care, marketing well in excess of available research, and legacy regulations now applied to telehealth, which leaves individuals vulnerable to inadequate online mental health resources in times of need (Essig, 2015a; Mahtta et al., 2021; Zeavin, 2022). These concerns span the array of telepsychology tools, including remote sessions via audio or video where reduced non-verbal cues and behaviors interfere with conversational rhythms and verbal fluency. This, in turn, impacts processes that are fundamental to rapport building, including social presence, empathy, and physical and emotional intimacy (Isaacs-Russell, 2015; Essig, 2015b; Roesler, 2017; Simpson, 2009).When it comes to online psychoeducational tools, they are limited in their inability to provide information that is personalized and clinically relevant to the user, and often vary in quality and accuracy (Barak & Grohol, 2011). Regarding interactive self-help interventions, low adherence and significant dropout rates are a common problem that prevents many individuals from experiencing the programs' benefits (Barak & Grohol, 2011; Lipschitz et al., 2022). Internet-based psychological interventions are also unable to accurately detect when an individual is in crisis or in need of alternative treatment services, which presents serious ethical and clinical challenges (Barak & Grohol, 2011; Calvo, Milne, Hussain, & Christensen, 2017; Pham, Nabizadeh, & Selek, 2022).

2. Can a chatbot provide therapy?

A separate and final group of behavioral intervention technologies is Artificial Intelligence (AI) based machines. This emerging behavioral intervention technology is used for promoting communication, social activity, guidance, education, day-to-day functioning, and entertainment (Pham et al., 2022; Shibata & Wada, 2011). Some have argued that AI-based chatbots may be able address problems with adherence and lack of personalization and interaction of other BITs, as they are uniquely well positioned in a middle ground between professionally provided psychotherapy and self-help apps due to the potential for a therapeutic chatbot to simulate aspects of therapy that are predictive of mental health benefits, including a sense of connection, alliance, and enlightenment (Thompson, 2018; Pham et al., 2022).

An early chatbot—originally designed to illustrate conversational computer access as opposed to batch processing by mimicking Rogerian psychotherapy—was ELIZA, an automated response system developed between 1964 and 1966 that turned users' statements into questions using Rogerian techniques of restatement and empathic reflection. To the surprise of the creator, users experienced a conversation that was both non-directive and client-centered (Cavanagh, Zack, Shapiro, & Wright, 2003; Thompson, 2018; Weizenbaum, 1966). The program's natural language system enabled individuals to use a teletype keyboard to communicate with the therapy-bot program using ordinary words, phrases, and sentence structure, thereby creating an experience of conversational interaction (Cavanagh et al., 2003; Mello & Souza, 2019; Pham et al., 2022; Weizenbaum, 1966).

In the decades that followed, AI bots were developed along the entire dimension of self-help to psychotherapy (Zeavin, 2022). While ELIZA used psychotherapy discourse to illustrate computerized natural language processing, SHRINK – a chatbot intended to simulate psychotherapeutic dialogue – represented a failed attempt to build a psychotherapist (Wilson, 2010; Zeavin, 2022). Other chatbots were more successful. Some offered low levels of mental health support and companionship to people with dementia, where they were found to help with affect regulation and decrease stress (Mordoch, Osterreicher, Guse, Roger, & Thompson, 2013; Pham et al., 2022). They were also found to help in decreasing feelings of isolation and depression in elderly patients when used as health care assistants; and promote social skills, engagement and attention in children suffering from autism spectrum disorders (Costescu, Vanderborght, & David, 2014; Fiske, Henningsen, & Buyx, 2019; Pham et al., 2022).

More recent technological advances in communication have influenced a new wave of AI machines that feature a wider array of psychotherapeutic techniques aimed at performing higher level interventions that were traditionally performed by psychotherapists and other (human) professionals. Some chatbots involve very focused interventions, for example addressing hallucinations and other treatment-resistant psychotic symptoms by providing patients who have schizophrenia with an opportunity to engage with voices and images of people in their lives via the bot (Fiske et al., 2019; Pham et al., 2022). Meanwhile, therapeutic chatbots across the self-help to psychotherapy dimension are increasingly being used by individuals with a wide variety of mental health issues, including those who may not meet criteria for a specific diagnosis, who use the chatbot's functionality to receive responsive and guided conversation, advice and strategies for coping with mental health difficulties (Fiske et al., 2019).

3. Woebot: research and application

Woebot, developed in 2017 to offer a structured, manualized cognitive-behavioral therapy for people experiencing depression, anxiety, or other mental health difficulties, was the first therapy chatbot to be tested in a randomized controlled (RCT) design (Fitzpatrick, Darcy, & Vierhile, 2017). Findings indicated that compared to people in a control condition who read psychoeducational materials about depression, individuals who used Woebot showed a decrease in symptoms of depression; the investigators concluded that Woebot and other similar conversational agents are a "feasible, engaging, and effective way to deliver CBT" (Fitzpatrick et al., 2017, p. 19e). Concluding that a relationship with Woebot can provide CBT implicates a level of care beyond self-help behavioral intervention technologies; it stakes a claim that Woebot is a psychotherapy provider. This raises questions about the evidence required to promote an AI-based chatbot as providing or delivering psychotherapy or mental health services. In psychotherapy research, best practices indicate that new treatment modalities must show efficacy in comparison to other treatment modalities. Claims of a treatment's effectiveness can then be made in relation to another treatment (often referred to as "treatment as usual"), which has been found to be effective in treating certain symptoms or diagnoses (Donker, Griffiths, Cuijpers, & Christensen, 2009). While the research on Woebot utilized a standard RCT model, it did not include a treatment-*as*-usual group, such as CBT delivered by a human psychotherapist. Conversely, given the body of research noting widespread benefits of self-help

improvement programs and applications (Anthony, 2003; Costescu et al., 2014; Fiske et al., 2019; Mordoch et al., 2013), the findings from Woebot seem to suggest it is on par with other BITs.

Relatedly, the decision to use a no-treatment control group RCT design to support a claim of Woebot's effectiveness is problematic; reading pamphlets about depression is not a known effective means of decreasing depressive symptoms. To the contrary, research suggests that passive psychoeducation, which is characterized by non-guided interventions that usually include educational materials and advice, tends to have weak effect sizes (Riper et al., 2009). On the other hand, active psychoeducation (for example, materials that describe and teach CBT) are more consistently found to have significant positive effects (Donker et al., 2009; He et al., 2022). Along with randomization which reduces external variability, using comprehensive control conditions helps to ensure internal validity, which strengthens the chances that the outcome can be attributed to the experimental treatment and not an alternative explanation (Mohr et al., 2009). When researchers utilize less comprehensive control conditions, such as the Woebot-RCT no-treatment condition their results are likely to show larger between-group effects and smaller within group effects (Mohr et al., 2009). Therefore, while the claim that Woebot demonstrated greater efficacy than the control group may be accurate, the implication that Woebot can effectively deliver an evidence-based intervention may be misleading. In order to substantiate this claim with any degree of confidence, Woebot would ideally be included in a clinical trial alongside conventional face-to-face psychotherapy approaches to treating mental health difficulties with demonstrated effectiveness (He et al., 2022). However, this has yet to be done. Since their original article, studies on Woebot have expanded to include adolescents with anxiety and depression (Nicol, Wang, Graham, Dodd, & Garbutt, 2022) and individuals with substance misuse (Prochaska et al., 2021). All of these have followed the same model of comparing people who use Woebot to people assigned to read passive psychoeducational materials. The exception to this was two studies that focused on women with post-partum depression, which compared Woebot to treatment as usual (Darcy et al., 2022; Ramachandran et al., 2020). Surprisingly, however, no description of what characterized treatment as usual is included in these papers, which severely limits what can be concluded.

A second set of concerns regarding the original study arise regarding the analyses and results themselves. First, the authors base their conclusions regarding the efficiency of Woebot with very limited evidence. They used missing data imputation (without explaining the method used, only that it was done in SPSS) to impute data for intent-to-treat participants. Subsequently, they found that, with this procedure, a significant effect was observed on one out of four of the outcome variables (the PHQ-9). They then used a Bonferroni correction to correct the p value obtained, which was corrected very close to the significance threshold ($p < .04$). Based on the non-significant effects observed for the other outcomes, it is quite plausible that the effect might not have been significant without the use of missing data imputation. The fact that the same analysis was not reported without missing data imputation therefore jeopardizes the very conclusions of the authors' study, or at the very least call for a replication.

Second, there was a lack of consistency in reporting the analyses conducted amongst all treatment conditions within the results. Changes in depression between baseline and follow-up were evaluated using repeated measures dependent t-tests and Cohen's *d* effect sizes for each separate item of the PHQ-9 (Fitzpatrick et al., 2017). Although the study indicated that a dependent *t*-test was conducted on PHQ-9 items, findings were not reported. Additionally, Cohen's *d* effect sizes for each item in the depression scale were reported for participants in the Woebot group, but not those in the control group. In the interest of assessing the specific benefits of Woebot verses the control group, the findings are incomplete as the analysis did not account for group differences and in turn, did not address the incremental benefit of Woebot. A repeated measures ANOVA for all outcome variables that compares participants in the two groups who completed both baseline and follow-up measures would be a more standard statistical design in such a study.

4. The current study

Given the high stakes that exist when people seek services to help with mental health problems and the vast sums of money being invested in telepsychology tools like Woebot (Delgone, 2022), it is critical that studies engage in thorough, careful examination of their effects, which can be used to disseminate clear information to the public. While existing literature is quite comprehensive in noting the benefits and pitfalls of using chatbots to address mental health difficulties, no study has demonstrated whether these chatbots offer benefits above and beyond the wide array of self-help behavioral intervention technologies that are available on the market. The current study thus aimed to engage in a rigorous examination of the effects of Woebot in an RCT with multiple comprehensive control conditions featuring active components. This design both allowed for replication of the initial findings, while also extending inquiry into how a chatbot that aims to offer psychotherapy performs in relation to self-administered behavioral intervention technologies in which individuals access technologies to address their own mental health needs. As such, the current study examined the efficacy of Woebot in relation to three other conditions: (1) psychoeducational materials regarding depression (passive control, identical to original study); (2) ELIZA, that computer program that mimics a Rogerian therapist (active control, intended to replicate the conversational component of Woebot); and (3) an interactive writing task (active control, intended to replicate the opportunity for self-expression offered by Woebot).

The inclusion of active controls aligns with recommendations of Donker et al. (2009), to use alternatives to passive psychoeducational materials for attention-placebo control groups, which, as described above, can obscure the true effect size in the intervention group. Utilizing active controls within this study was expected to eliminate differences amongst the changes in anxiety, depression, and positive/negative affect between the intervention group and the active control groups. On the other hand, differences in levels of anxiety, depression and positive/negative affect were expected when comparing the active interventions to the passive control group. Therefore, it was expected that levels of anxiety, depression, and negative affect would decrease, and positive affect would increase, among participants in the three active intervention groups (Woebot, ELIZA, and Daylio), with equivalent effect sizes. Secondly, no significant change in any mental health outcomes were expected for participants in the passive informational control group.

5. Methods

5.1. Participants and procedures

Participants were recruited on Facebook's social media platform between May and June 2021 using advertisements targeting college and graduate students between the ages of 18–29 who identify with symptoms of anxiety and/or depression. Consistent with procedures from Woebot's original study (Fitzpatrick et al. (2017), eligible individuals were presented with an opportunity to receive up to $20 and receive technology delivered self-help materials to promote mental health. In addition to age and student status, participants were required to have an android or an iPhone to access intervention materials. After an initial screener to determine eligibility, eligible individuals took a 10-min survey that included all measures, which culminated in being randomized using a computer algorithm that automatically assigned participants to one of four intervention groups upon survey completion. Participants were then sent an email from an email address developed for this study containing instructional information pertaining to their intervention group, and a10-dollar amazon gift. This process continued until 120 participants were signed up and randomly assigned into one of

the four intervention groups, to stay consistent with the original study's inclusion of 30 participants per group.

All participants received an email with instructions to access their assigned intervention, followed by daily reminders to engage with the intervention daily for two weeks. Participants using Woebot downloaded an app, where they could chat with the bot at any time. Participants using ELIZA accessed the program through a website messaging app, where they could chat with the bot at any time. Daylio was downloaded via an app, where participants were invited to keep a daily interactive journal (see Chaudhry, 2016). Participants in the Psychoeducation group received an email instructing them to read the materials used in the original Woebot study, entitled "Depression in College Students" (National Institute of Mental Health, 2017).

After two weeks, participants received an email containing a link to a second 15-min survey, which included the same measures as the baseline study. Upon survey completion, participants received a second 10-dollar amazon gift card via email.

5.2. Measures

Demographic information, including age, sex, gender, sexual orientation, and racial background, was collected at baseline. All other measures were collected at both baseline and follow-up. All outcome measures match the original Woebot study.

Anxiety was measured by the Generalized Anxiety Disorder 7-item scale (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006), a frequently used and well-validated measure that evaluates the severity and frequency of anxious thoughts and behaviors over the course of two weeks (sample items include feeling nervous, anxious, or on edge). Items are assessed on a scale from 0 (not at all) to 3 (nearly every day). After summing all items, scores that are greater than 10 indicate moderate anxiety, while scores greater than 15 indicate severe anxiety (Spitzer et al., 2006). It demonstrated very good to excellent reliability ($\alpha = 0.91$ at baseline) and follow-up ($\alpha = 0.89$ follow-up).

Depression was measured by Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001), a frequently used and well-validated measure that assesses the frequency and severity of depressive symptoms over the course of two weeks (sample items include feeling down, depressed, or hopeless). The items are scored on a 0 (not at all) to 3 (nearly every day) scale, with total scores from 0 to 5 indicating no symptoms of depression, scores of 5–9 indicating mild symptoms, scores 10–14 indicating moderate symptoms, scores 15–20 indicating moderately severe symptoms, and scores greater than 20 indicating severe depression (Kroenke et al., 2001). It demonstrated very good reliability ($\alpha = 0.88$ at baseline) and ($\alpha = 0.89$ follow-up).

Positive and negative affect were measured by the Positive and Negative Affect Schedule (PANAS; Watson et al., 19) based on identification with ten words connoting positive affect (sample items include interested, excited, and strong) and ten words connoting negative affect (sample items include distressed, upset, and guilty). Individuals assign a score of the extent to which they feel each emotion (from 1 = Very Slightly or Not at All to 5 = Extremely). Scores range from 10 to 50, with higher scores indicating greater positive or negative affect (Watson, Clark, & Tellegan, 1988). In this study, the PANAS showed very good reliability at baseline and follow-up for positive affect ($\alpha = 0.89$ at both time points) and negative affect ($\alpha = 0.90$ at baseline and 0.89 at follow-up).

5.3. Statistical analysis

The preliminary analysis included frequencies on demographic variables between all study variables to identify any abnormalities within the data. Randomization and attrition checks were conducted using a One-Way analysis of variance (ANOVA) and independent samples chi-square tests to evaluate whether there were significant differences in demographics, baseline measures of anxiety, depression, and positive and negative affect between participants who were enrolled in the Woebot, ELIZA, Daylio and psychoeducation groups.

5.4. Main analysis

Two-Way Mixed ANOVA models with a repeated measures factor (time), a group factor (treatment group) and the interaction between the two were used to test for differences between groups in the evolution over time of the outcome variables. Type III tests were used to test all effects. The test of the interaction effect between time and the group variable was particularly regarded, as it tests whether groups had statistically significant differences in symptom reduction and negative affect reduction over time – in other words, it tests whether the efficacy of the treatments differed.

Similar models were also run in a multilevel (i.e., hierarchical, mixed-effects) modeling framework, using the popular package "lme4" in R. We fitted the same models as described above, with the addition of a random intercept term. This procedure allowed us to account for the cases having random (as opposed to fixed) baseline levels, and therefore more accurately represents the research design. More complex models with random slopes were attempted, but could not be estimated, due to low group sizes. Type III F-tests with Satterthwaite's degrees of freedom method were used to test effects, as implemented in the R package `lmerTest".

5.5. Secondary analysis

Paired samples t-tests were used to evaluate intervention effects between baseline and post intervention within each group specifically. This allowed for specific conclusions to be made about the efficacy of each intervention group following a two-week intervention period. Cohen's *d* was calculated to indicate the effect size of the observed changes.

6. Results

6.1. Preliminary analysis

Of the 191 respondents registered for the study, 54 respondents did not progress through the study due to ineligibility or failure to complete the baseline survey. Of the remaining 137 respondents who met eligibility criteria for the study, 17 respondents were eliminated due to duplicate submissions, failure to provide a verifiable email address, or identification as a potential bot by the survey's software perusal. This resulted in a sample of 120 participants, 30 of whom were randomized into each of the four groups.

After the two-week intervention period, 77 participants submitted post intervention surveys. Of these, 12 submissions were eliminated from the data due to incomplete submissions or misreported treatment groups. The remaining sample was 65 participants, split across the four groups: Woebot (N = 18); ELIZA (N = 18); Daylio (N = 15); Psychoeducation (N = 14). Data collected from the sample of N = 65 was utilized for the final analysis.

6.1.1. Attrition

Out of the 120 randomized participants, 54% (65/120) provided complete data for the post intervention survey, representing an overall attrition rate of 46%. More White participants dropped out of the study before the second survey than non-White participants – X^2 (1, N = 120) = 10.79, $p = .001$. There were no significant differences related to attrition on the following variables: Hispanic/Latinx – X^2 (1, N = 120)= 0.46, $p = .5$, sex – X^2 (1, N = 120) = 0.08, $p = .777$, gender – X^2 (2, N = 120) = 1.7, $p = .428$, and sexual orientation – X^2 (1, N = 120) = 1.7, $p = .193$. No differences were found in anxiety – $F(3,118) = 0.02$, $p = .899$, depression – $F(3,118) = 1.21$, $p = .273$, positive affect – $F(3,118) = 0.25$, $p = .620$, or negative affect – $F(3,118) = 0.01$, $p = .916$ between

individuals who did and did not complete the follow-up survey. There were no significant age differences – $F(1,118) = 0.06$, $p = .810$ between participants who completed the study versus those who did not.

6.1.2. Randomization

No significant differences were found amongst the four treatment groups at baseline in terms of race – X^2 (3, N = 120) = 6.93, $p = .074$, ethnicity (Hispanic/Latinx) – X^2 (3, N = 120) = 2.18, $p = .536$, gender – X^2 (6, N = 120) = 5.39, $p = .496$, or sexual orientation – X^2 (3, N = 120) = 4.09, $p = .252$. There were no significant age differences amongst the groups – F (3,116) = 1.9, $p = .134$. The same was observed for anxiety – $F(3,116) = 0.536$, $p = .66$, depression – $F(3,116) = 0.54$, $p = .659$, positive affect – $F(3,116) = 0.51$, $p = .680$ and negative affect – $F(3,116) = 1.37$, $p = .255$.

6.1.3. Completer randomization

A one-way ANOVA was conducted on participants who completed the study (N = 65) to assess baseline differences in clinical measures between groups to evaluate appropriate randomization. There was no significant difference between groups in baseline measures of anxiety – $F(3,61) = 0.14$, p = .935, depression – $F(3,61) = 0.02$, p = .995, as well as positive affect – $F(3,61) = 1.68$, p = .180 and negative affect – $F(3,61) = 0.11$, p = .955 measures of the PANAS scale.

6.2. Participant demographics

As found in Table 1, participants included 65 individuals (Mage = 22.97 years; SD = 3.02) who completed both baseline measures and post intervention measures for their assigned intervention. More than two-thirds of the participants identified as female (70.8%), with the remaining participants identifying as male (15.4%) or other (13.8%), which included as non-binary, trans or intersex. More than half of the participants were White (56.9%), with 23.1% Asian, 10.8% Black/African American, 3.1% American Indian/Alaska Native, and 6.2% Other, which included Biracial, Middle Eastern and Caribbean Indian. Participants were mostly non-Hispanic (92.3%).

6.3. Main analysis

As seen in Table 2, based on the Mixed Two-Way ANOVA models (with fixed effects only), the effect of time, treatment group and their interaction were studied. Results indicated that there was no two-way interaction between time and treatment group regarding anxiety – $F(3,61) = 1.31$, $p = .279$, depression – $F(3,61) = 0.53$, $p = .666$, positive affect – $F(3,61) = 1.22$, $p = .311$, and negative affect – $F(3,61) = 0.67$, $p = .570$). There was however a main effect of time – in the direction of improving health outcomes – on all four outcome variables: anxiety – $F(1,61) = 21.61$, $p < .001$, depression – $F(1,61) = 25.58$, $p < .001$, positive affect – $F(1,61) = 9.86$, $p = .003$, and negative affect – $F(1,61) = 21.61$, $p < .001$.

The same models with random intercept indicated nearly identical findings with a non-significant time-by-group interaction on anxiety – $F(3,61) = 1.31$, $p = .279$, depression – $F(3,61) = 0.53$, $p = .666$, positive affect – $F(3,61) = 1.22$, $p = .311$, and negative affect – $F(3,61) = 0.68$, $p = .570$. Similar to the results obtained with the fixed-effects model, significant main effects of time were observed regarding anxiety – $F(1,61) = 21.61$, $p < .001$, depression – $F(1,61) = 25.78$, $p < .001$, positive affect – $F(1,61) = 9.86$, $p = .003$, and negative affect – $F(1,61) = 16.4$, $p < .001$ pointing to the overall efficacy of the interventions. The estimated marginal means of this model, along with 95% confidence intervals and individual trajectories, are showed in Figures 1, 2, 3 and 4

6.4. Secondary analysis

To more specifically interpret the impact of each intervention, paired samples t-tests were conducted with the outcome measures in all four intervention groups.

6.4.1. Anxiety

The results indicated a significant decrease between baseline anxiety levels (T1) and post intervention anxiety levels (T2) amongst participants in the Woebot group – $t(17) = 3.66$, $p = .002$, $d = 0.863$ and the ELIZA group – $t(17) = 3.94$, p = .001, $d = 0.928$. A marginally significant decrease was found between baseline anxiety levels (T1) and post intervention anxiety levels (T2) amongst participants in the Daylio group – $t(14) = 2.1$, $p = .055$, $d = 0.524$. No significant differences were found between baseline anxiety levels (T1) and post intervention anxiety levels (T2) amongst participants in the Psychoeducation group – $t(13) = 0.58$, $p = .575$, $d = 0.154$.

6.4.2. Depression

The results indicated a significant decrease between baseline depression levels (T1) and follow-up depression levels (T2) amongst participants in the ELIZA group – $t(17) = 4.39$, $p < .001$, $d = 1.035$ and the Daylio group – $t(14) = 2.21$, $p = .045$, $d = 0.570$. The results indicated a marginally significant difference between baseline depression levels (T1) and post intervention depression levels (T2) amongst participants in the Woebot group – $t(17) = 1.91$, $p = .073$, $d = 0.450$ and the Psychoeducation group – $t(13) = 2.08$, $p = .058$, $d = 0.555$.

6.4.3. Positive affect

The results indicated a significant increase between baseline positive affect levels (T1) and follow-up positive affect levels (T2) amongst participants in the ELIZA group – $t(17) = -3.49$, p = .003, $d = -0.822$ and a marginally significant increase amongst participants in the Psychoeducation group – $t(13) = -2.05$, $p = .061$, $d = -0.583$. The results indicated that there was no significant difference between baseline positive affect levels (T1) and follow-up positive affect levels (T2) amongst participants in the Woebot group – $t(17) = -0.01$, $p = .979$, $d = -0.006$ and the Daylio group – $t(14) = -1.71$, $p = .11$, $d = -0.441$.

6.4.4. Negative affect

The results indicated a significant decrease between baseline negative affect levels (T1) and follow-up negative affect levels (T2) amongst participants in the ELIZA group – $t(17) = 3.34$, $p = .004$, $d = 0.788$ and the Daylio group – $t(14) = 2.38$, $p = .032$, $d = 0.615$. There was no significant difference between baseline negative affect levels (T1) and follow-up negative affect levels (T2) amongst participants in the Woebot group – $t(17) = 1.56$, p = .138, $d = 0.367$ and the Psychoeducation group – $t(13) = 1.10$, $p = .29$, $d = 0.295$.

7. Discussion

The findings from this study provided partial support for the hypotheses that Woebot would not differ from other app-based (non-psychotherapy-identified) interventions in leading to improvements in mental health, and that all three active interventions would be superior to a passive informational control group. The main analysis failed to detect differences between any of the four treatment conditions in improving symptoms of depression, anxiety, and positive/negative affect, likely due to a lack of power in looking at all four groups simultaneously. However, secondary analyses using paired samples t-tests provided additional information about where significant improvements were found for each intervention separately, along with their effect sizes. Findings indicated most robust effect sizes for ELIZA users, followed by those using Daylio, then Woebot and finally Psychoeducation. Participants who used ELIZA experienced significant improvements in all four outcome areas, with large effect sizes for anxiety, depression, and positive affect, and a medium effect size for negative affect. This was followed by Daylio, where participants reported significant decreases in depression and negative affect (with medium effect sizes) and a marginally significant decrease in anxiety,

also with a medium effect size. Next, participants who used Woebot reported a significant decrease in anxiety with a large effect size and a marginally significant decrease in depression with a small effect size. Finally, participants who received psychoeducation had a marginally significant reduction in depression and a marginally significant increase in positive affect, both with medium effect sizes.

This study did not replicate Fitzpatrick et al. (2017)'s findings regarding Woebot's effectiveness over a passive informational control in alleviating depressive symptoms, as there were marginally significant decreases in depression levels for participants in both the Woebot and psychoeducation groups. In the current study, benefits to Woebot users were found for anxiety rather than depressive symptoms; however, this improvement was on par with ELIZA, a non-psychotherapy conversational bot. Further, both ELIZA and Daylio—included in this study as per recommendations by Mohr et al. (2009) to increase validity by testing the active intervention to active control groups that were meant to exemplify the expressive and conversational elements of Woebot—were found to lead to improvements in symptoms that were equivalent to or greater than the benefits associated with Woebot.

Overall, findings support previous research illustrating the benefits of behavioral intervention technologies. The benefits experienced by participants in the journaling group support previous studies finding that daily journaling via an app promotes creativity and resilience and improves mood by affording individuals the opportunity to further observe and reflect on their thoughts and behaviors (Jeong & Breazeal, 2016; Oduntan et al., 2022). Interestingly, of all three interventions, ELIZA was found to be most broadly associated with symptom improvement. Previous research has found that chatbots can be beneficial for both clinical and community-based samples in addressing mental health symptoms (Fiske et al., 2019). ELIZA's creator, Joseph Weizenbaum, described that programs such as ELIZA provide individuals with a sense of sympathy and feelings of being understood. He further described that people often perceived ELIZA's feedback to be words of wisdom, which may have contributed to a sense of emotional connection towards the bot (Thompson, 2018). Notably, and in line with the outcome of the current study, Weizenbaum shared his surprise about the intensity of users' response to ELIZA, which propelled concerns that the rise of chatbots would engender therapeutic structures that are conditional and/or automatic in nature (Thompson, 2018). He expressed a conviction that bots should not be used to replace psychotherapy due to the "interpersonal respect, understanding, and love" that cannot be provided with computerized psychotherapy (1976, p. 269).

7.1. Clinical & societal implications

Along the lines of Weizenbaum's comments, the current study echoes growing concerns that the affordances of digital technology (convenience, ease and cost) are leading to complacency and a reduction in the quality of what is considered mental health treatment (Zeavin, 2022). Based on only one study conducted by its creators and no known attempts at replication, Woebot is currently being marketed on its site as "highly researched … mental health care." Mental health care, defined as "services devoted to the treatment of mental illnesses and the improvement of mental health in people with mental disorders or problems" (Collins, 2022, para. 1), has typically referred to a limited arena of professional care governed by a set of ethics and guidelines for practice, that is clearly distinguishable from the many actions and behaviors that people can take in order to improve their mood, such as exercising or journaling (both of which have been found to significantly improve mood symptoms; see Brodwin et al., 2016). As described in the literature review, digital technology has significantly increased the array and complexity of self-help programs, with many benefits for personal growth and overall well-being (Simmons et al., 2016). Mental health applications such as Woebot provide individuals with alternatives to traditional talk therapy to address mental health issues at no cost (Dinneen, 2020), which makes them extremely attractive (Conley et al., 2022). Against this backdrop, it is arguably even more critical for there to be clarity around the standards of what can be considered mental health care.

For individuals looking for psychotherapy, the claim that Woebot offers a mental health treatment is then augmented by the focus on Woebot's human qualities. Woebot was originally created as a conversational agent that is programed to deliver messages to users that are empathic and personalized while also remaining transparent in its presentation as an artificial agent in effort to avoid representing itself as an artificial agent trying to pass as a human (Darcy, Daniels, Salinger, Wicks, & Robinson, 2021). However, Woebot's website particularly features findings of users expressing that talking to Woebot feels like talking to a human being who shows concern, as well as research findings that Woebot users' report feelings of human-level bonding after interacting with Woebot for five days (Darcy et al., 2021). Woebot makes statements about its own humanity, such as "I'm like a wise little person you can consult with during difficult times, and not so difficult times," then later following up with "I'm not a human," and "But in a way, I'm still a person" (Dinneen, 2020). This suggests that despite the creators' stated intention to highlight Woebot's artificial nature, Woebot has been designed to emphasize its humanness and availability to offer therapeutic care during hard times. Moreover, Woebot's developers have conducted research to demonstrate that people develop a bond to Woebot that is commensurate to the bond that people develop to their human therapists (Darcy et al., 2021).

People seeking help for mental health difficulties are often in a vulnerable position, and it is common to turn to the Internet for information about treatment. While the Internet is, of course, filled with various claims about products of all kinds, with varying degrees of accuracy, the stakes are particularly high when the marketing of mental health treatment is likely to influence people's capacity and decisions to seek help for their mental health symptoms (Conley et al., 2022). Rates of anxiety and depression have increased, as have challenges to accessing traditional face-to-face therapy, due to factors including insufficient insurance coverage and lack of available providers (Zeavin, 2022). This calls for ethical and responsible dissemination of information about the services that do exist based on research held to rigorous standards.

The procedures that would safeguard these goals have already been well-established. First, there is a pressing need for replication of any mental health improvements associated with new telepsychology programs with large and diverse samples of community members and clinical populations (Kretzschmar et al., 2019). Secondly, these interventions must be tested against treatment-*as*-usual (e.g., CBT administered in person) to support any statements of efficacy that equalize the bots with mental health treatment (e.g., Lucas, Gratch, King, & Morency, 2014; Rizvi, Dimeff, Skutch, Carroll, & Linehan, 2011), and to demonstrate the magnitude of effects in relation to standard treatment (Kretzschmar et al., 2019). These measures are crucial for preventing the messaging about therapy chatbots from being taken out of context. Younger people may be particularly vulnerable to messaging via the Internet about the benefits of turning to chatbots to improve mental health, especially as the need for quality treatment far surpasses its availability (Conley et al., 2022). Relatedly, many of the reasons that might incline young people towards chatbots are driven by their reluctance to access mental health care, including the wish to avoid stigma about accessing services, a desire for self-reliance when coping with emotional problems, greater comfort when using text than communicating in person, and a reluctance to trust humans with sensitive and private information (Conley et al., 2022; Lucas et al., 2014). When these motivations are at play, use of chatbots can be further undermining to mental health and well-being and may even give rise to the underlying mental health difficulties that are bringing people to therapy in the first place—with one example being an exacerbation of alienation or isolation (Kretzschmar et al., 2019). There are also serious concerns that arise related to privacy (for example, Woebot is available

through Facebook Messenger, which is attached to people's identities), and to safety (including the pitfalls of people becoming overly reliant on therapy bots and/or relying on the bot in emergency situations that the bot cannot deal with). These issues, alongside the understanding that chatbots do not replicate the authenticity and individualized attention that comes from human interactions, raise serious concerns about replacing traditional therapy with therapeutic chatbots (Kretzschmar et al., 2019).

Part of the issue has to do with a lack of precision in the language used to describe AI bots, which are interchangeably labeled as therapy bots and social bots. This is the case despite their significantly divergent operating systems and functions: While social chatbots tend to be limited to providing informational support, therapy chatbots attend to users' emotional needs by generating complex responses which allow individuals to take the lead on the conversation (Ahmed et al., 2023; Wang, Mujib, Williams, Demiris, & Huh-Yoo, 2021). The issues of terminology in AI bots mirror what is happening on a large scale, as our language for what constitutes being 'in therapy' has become increasingly vague and all-encompassing, essentially collapsing the use of an app with formal psychotherapy sessions conducted by a trained, licensed professional (Zeavin, 2022).

Finally, the current study's inclusion of multiple interventions that represented the different elements of Woebot is in line with a movement in psychotherapy outcome research that calls for a focus on studying the mechanisms that help people in therapy, rather than solely focusing on outcomes such as symptom reduction (Norcross & Wampold, 2011; Wachtel, 2010). While this study's findings emphasize the importance of retaining a clear distinction between BITs and mental health treatment, it does suggest that journaling might be particularly indicated for individuals experiencing depressive symptoms, and that the conversational nature of a chatbot can be beneficial for both depression and anxiety. It is notable that larger effect sizes across outcomes were found for ELIZA than Woebot, despite Woebot's inclusion of advanced technologies to offer "smart" responses. It is conceivable that simple, Rogerian-inspired mirroring of what an individual has said may seem more believable than a "smart" bot, given the participants' awareness of the non-human nature of the "therapist." Further research could study the mechanisms that underlie different types of chatbots, and examine the notion supported by the current study that individuals who use self-help apps are able to engage in a depth of work on the self that leads to improved mental health and well-being, without any additional "smart" or human-like components. Such benefits can augment the benefits of therapy but are clearly distinct from professional therapeutic services.

Currently, there are thousands of mental health applications that are commercially available to the public but which have not gone through the rigorous development and testing procedures expected in applied research studies. This has led to an ever-expanding gap between technology driven mental health tools and the consistent application of evidence-based principles in supporting mental health (Conley et al., 2022). Meanwhile, research on mental health applications is increasingly calling attention to various risks associated with its usage, including the lack of evidence-based approaches (Koh, Tng, & Hartanto, 2022). A recent review of mobile interventions for youth found an inverse relationship between study quality and effect size, suggesting that better designed studies are actually *less* likely to support the efficacy of these interventions (Conley et al., 2022). Moreover, amongst higher-quality studies, the effect sizes in studies that included clinical comparison groups (as opposed to no-intervention control groups) were no longer statistically significant. (Conley et al., 2022).

The current findings uphold and extend these concerns, and support the notion that use of a no-treatment control group study to market a clinical service should no longer be considered an acceptable precondition to putting a new technological intervention on the market. The continuation of this practice places undue risk on vulnerable individuals seeking help for mental health difficulties. It may also call into question the efficacy of all technology driven mental health tools, if there is no quality control when it comes to studying their effects. These points have been emphasized by findings that mental health applications that are driven by the market put a spotlight on engaging and appealing qualities rather than prioritizing research driven by evidence-based practices and study design (Conley et al., 2022).

7.2. Limitations & future research

Although participants received daily reminders to interact with their assigned intervention, there was no way to accurately measure how often participants interacted with their assigned intervention. An umbrella review highlighted the prevalence of high attrition rates and low engagement in studies of mental health applications across six separate reviews (Koh et al., 2022). Relatedly, the attrition rate of 46% poses a threat to validity. Previous research suggests the high attrition can be attributed to the fully online setup of the current study, as attrition and poor treatment adherence are two issues commonly found in randomized control trial studies using smartphone interventions without any requirement to speak to a research team member by phone or in-person (Linardon & Fuller-Tyszkiewicz, 2020). Moreover, in the current study, the combination of high attrition and a relatively small sample size may have contributed to a lack of statistical power and resulting failure to detect differences in the analysis that included all four groups. Future studies should utilize larger samples to ensure statistical power and account for higher attrition rates in online studies.

Studies using monetary compensation (as was the case in both the current study and in Fitzpatrick et al., 2017) may reduce attrition threats at the cost of inflating treatment outcomes (Linardon & Fuller-Tyszkiewicz, 2020). However, as participant engagement with the interventions may have been related to compensation, this study is further limited by the difficulty of generalizing utilization and engagement to a naturalistic (non-research) setting. RCT studies on mental health applications have been shown to have larger effect sizes when their follow-up length is between seven and 11 weeks relative to those with a follow-up between two and six weeks (Linardon, Cuijpers, Carlbring, Messer, & Fuller-Tyszkiewicz, 2019). There are therefore limitations surrounding the length of treatment within this study. The importance in assessing the long-term effectiveness of technology driven mental health tools have also been noted (Economides et al., 2019; Rathbone, Clarry, & Prescott, 2017). It is unknown whether improvements were maintained after the study as long-term effects were not assessed. It is also important to consider the possibility that both passive and active groups may have experienced placebo effects when reporting their symptoms during follow-up.

Finally, the current findings should be interpreted only in relation to providing cognitive behavioral interventions with young adults experiencing symptoms of depression and anxiety regardless of symptomatic etiology (and not so severe as to preclude study completion). When researching chatbots designed to treat the generalized symptoms of depression and anxiety, it may be both more challenging and more urgent to identify the specific mechanisms of change being targeted when implementing these new interventions. Research on therapeutic chatbots that address symptoms experienced by individuals with autism spectrum disorder (ASD), schizophrenia, and dementia that was more directed towards specific symptoms of those disorders (Pham et al., 2022) provides an example of this approach.

8. Conclusion

The current study examined the efficacy of Woebot, which is marketed as a therapy bot that delivers CBT interventions, against two active interventions—ELIZA, a basic (non-"smart") conversational bot and a daily journaling app—and one passive control group where people read psychoeducational materials. While a repeated measures ANOVA conducted on all four groups simultaneously failed to detect any differences

between groups, paired samples t-tests examining differences in mental health outcomes before and after each intervention was administered showed that engagement with ELIZA and the daily journaling app resulted in decreases in depression of larger effect sizes than Woebot, and reductions in depression were equivalent between those who used Woebot and people in the passive psychoeducational control group. Participants using Woebot and ELIZA experienced equivalent decreases in anxiety symptoms. These findings underscore that Woebot performs on par with other behavioral intervention technologies directly challenging the notion that a conversational agent can be considered mental health treatment. Conversely, marketing a bot as therapy presents significant risks to individuals suffering mental health difficulties and looking for treatment.

When researching tools designed for symptom reduction and mental health support, these findings support the importance of an applied clinical approach of comparing the new to established treatments, rather than limiting the study to a basic science "better than nothing" approach. RCT studies that compare BITs to waitlist or psychoeducational controls that may contribute to basic science also possess significant risk of misuse. Such research has a high likelihood of being used to drive misinformation about the efficacy of smartphone applications to treat mental health problems, such as depression, anxiety, and substance use. As the popularity and scope of mental health technologies grow, developers of technology-driven mental health tools are economically incentivized to conduct research aimed at helping to market their interventions by providing "evidence" of their viability and efficacy. The "better than nothing" RCT is the tool of choice for this purpose. Rather, what is needed, and what is common in applied clinical research, is research that demonstrates the efficacy of a new intervention in relation to other evidence-based interventions and not just to no-treatment or psychoeducational controls. Also necessary is research that can identify the underlying mechanisms that contribute to whatever comparative efficacy is demonstrated. Such research can help illuminate our understanding of the tools that are most useful in mental and behavioral healthcare.

CrediT authorship contribution statement

Laura Eltahawy: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, data analysis. **Todd Essig:** Conceptualization, Writing – review & editing. **Nils Myszkowski:** Formal analysis. **Leora Trub:** Conceptualization, Writing – original draft, Writing – review & editing, data analysis.

Declaration of competing interest

The authors have no conflict of interest to disclose.

Appendix

Table 1
Demographic characteristics participant sample (N = 65)

	Number	Percent
Gender		
Male	10	15.4
Female	46	70.8
Other (Non-binary, Trans, Intersex)	9	13.8
Race		
White	37	56.9
Black/African American	7	10.8
American Indian/Alaska Native	2	3.1
Asian	15	23.1
Other (Biracial, Middle Eastern, Caribbean Indian)	4	6.2
Ethnicity		
Hispanic/Latinx	5	7.7
Sexual Orientation		
Heterosexual	25	38.5
Gay/Lesbian	5	7.7
Bisexual	19	29.2
Queer	13	20
Other (Asexual)	3	4.6
	M	*SD*
Age	22.97	3.02

Table 2
Mixed ANOVA for time and treatment condition

Measure	Predictor	Sum of Squares	df	Mean Square	F	p	partial η2
GAD-7 (Anxiety)	Time	226.92	1	226.92	21.61	<.001	.262
	Time*TC	41.3	3	13.77	1.31	.279	.061
	Error	640.47	61	10.5			
PHQ-9 (Depression)	Time	278.32	1	278.32	25.58	<.001	.295
	Time*TC	17.12	3	5.73	.527	.666	.025
	Error	663.78	61	10.88			
PANAS (Positive Affect)	Time	258.5	1	258.5	9.86	.003	.139
	Time*TC	95.71	3	31.91	1.22	.311	.056
	Error	1600.05	61	26.23			
PANAS (Negative Affect)	Time	569.6	1	569.6	16.4	<.001	.212
	Time*TC	70.48	3	23.49	.677	.570	.032
	Error	2118.24	61	34.73			

Note. Significant at the $p < .05$ level.

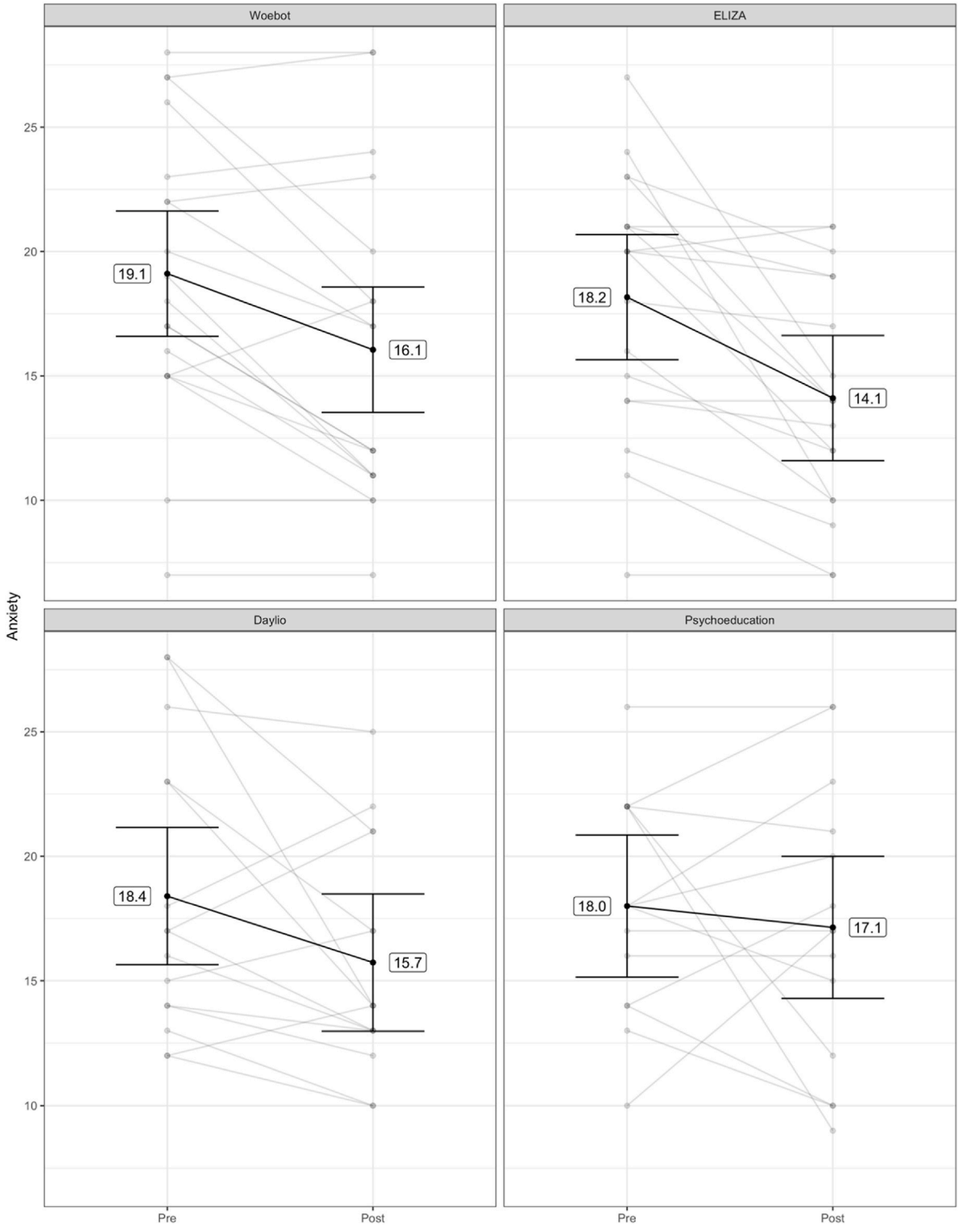


Fig. 1. Estimated marginal means and individual trajectories (Mixed Effects Model) of anxiety

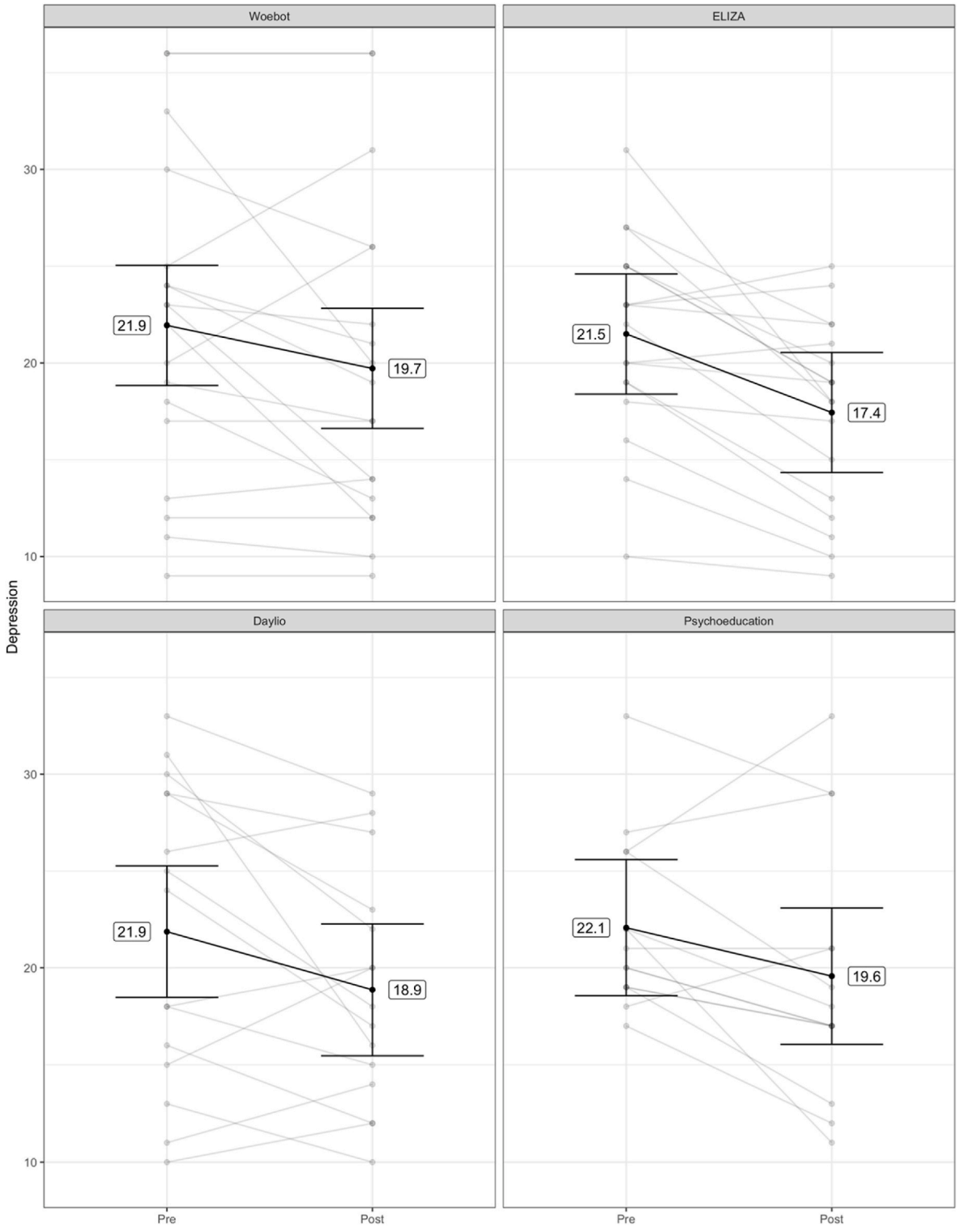


Fig. 2. Estimated marginal means and individual trajectories (Mixed Effects Model) of depression

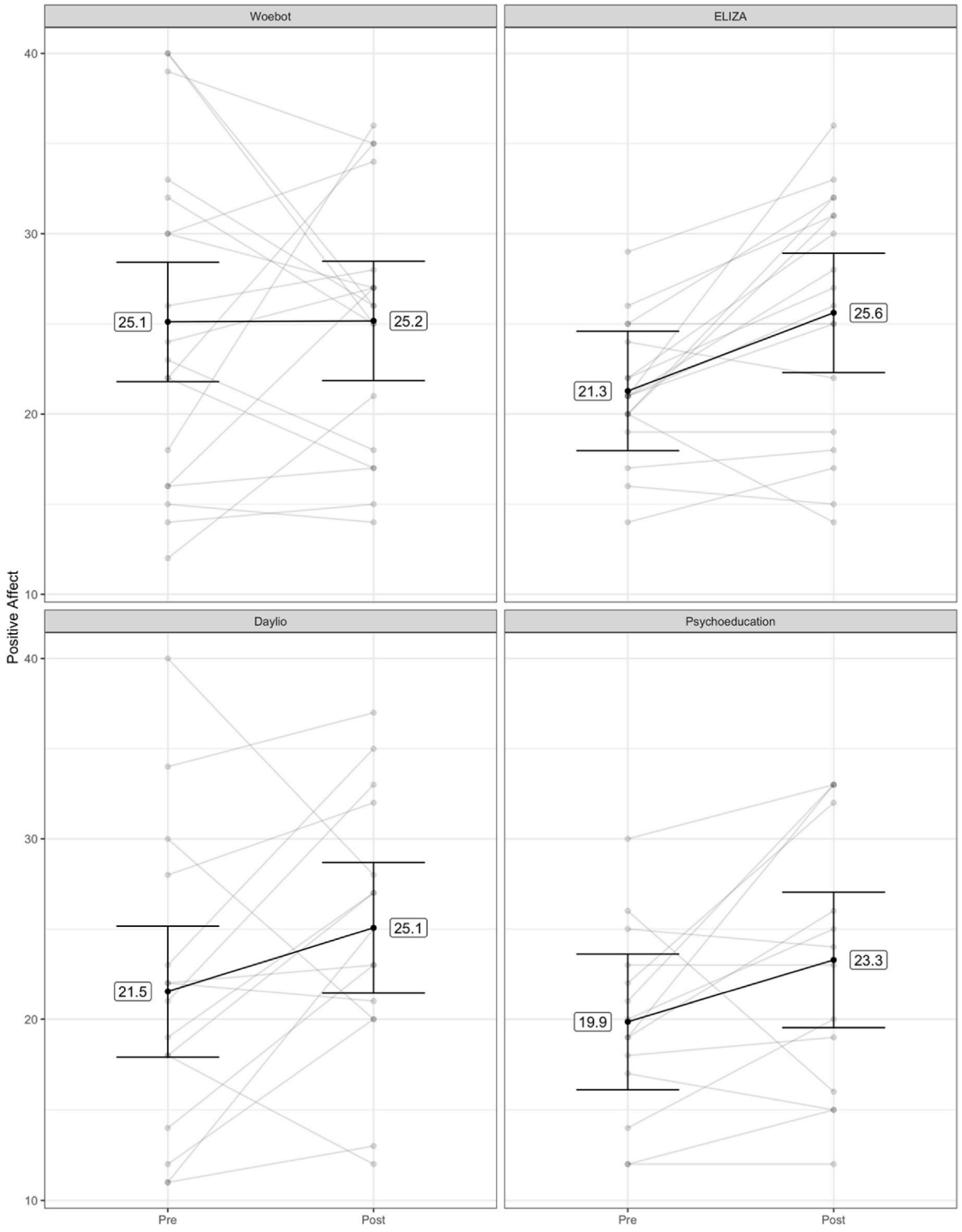


Fig. 3. Estimated marginal means and individual trajectories (Mixed Effects Model) of positive affect

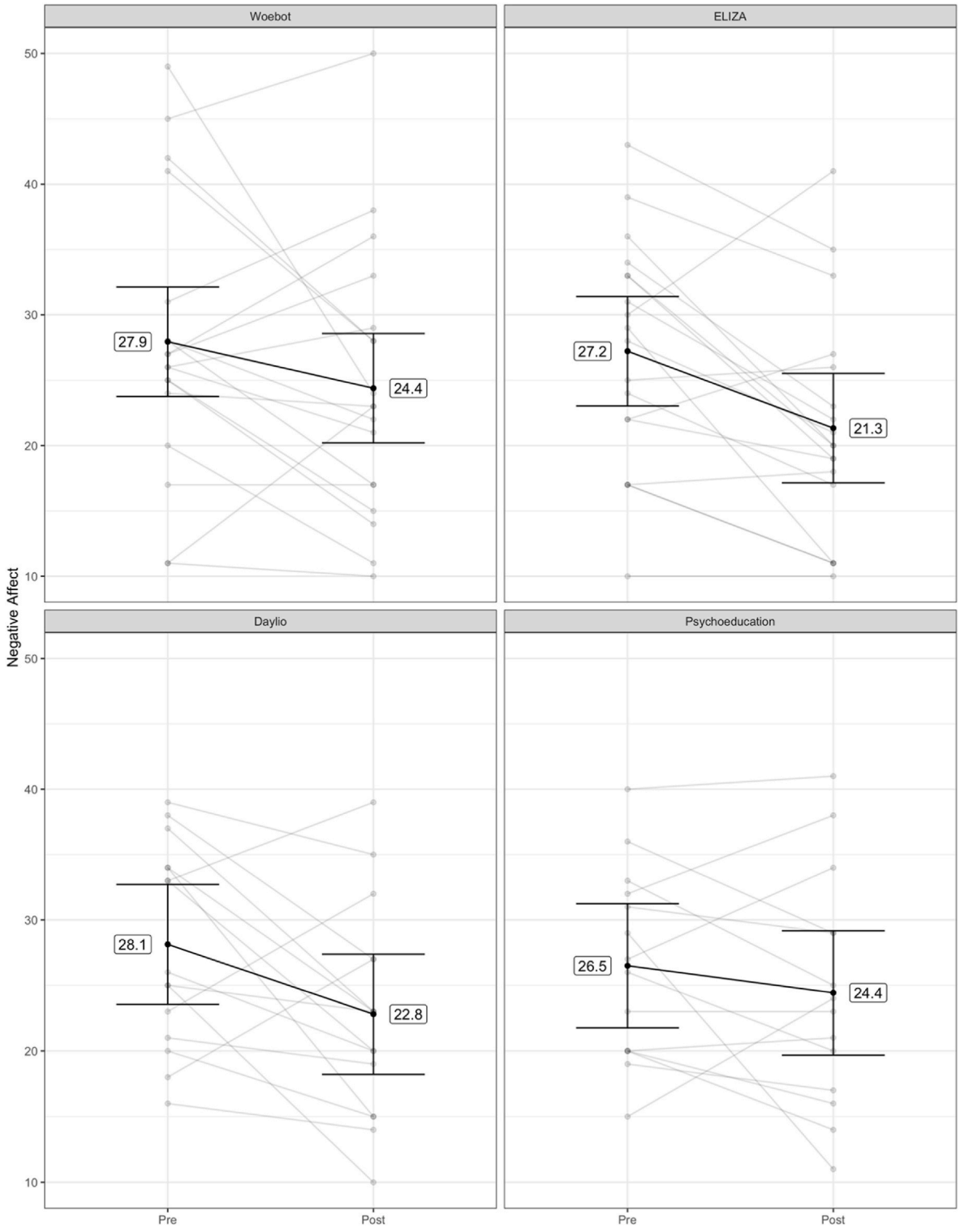


Fig. 4. Estimated marginal means and individual trajectories (Mixed Effects Model) of anxiety

References

Ahmed, A., Hassan, A., Aziz, S., Abd-Alrazaq, A. A., Ali, N., Alzubaidi, M., … Househ, M. (2023). Chatbot features for anxiety and depression: A scoping review. *Health Informatics Journal, 29*(1), 1376. https://doi.org/10.1177/14604582221146719

Anthony, K. (2003). The use and role of technology in counselling and psychotherapy. In S. Gross, & K. Anthony (Eds.), *Technology in counselling and psychotherapy: A practitioner's guide* (pp. 13–35). Macmillan Education UK.

Barak, A., & Grohol, J. M. (2011). Current and future trends in internet-supported mental health interventions. *Journal of Technology in Human Services, 29*(3), 155–196.

August 4 Brodwin, E., Orwig, J., & Spector, D. (2016). Here are 25 habits that psychologists have linked with happiness. *Business Insider* https://www.businessinsider.com/simple-ways-to-improve-your-mood-according-to-psychologists.

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

Cao, J., Zhang, G., & Liu, D. (2022). The impact of using mHealth apps on improving public health satisfaction during the COVID-19 pandemic: A digital content value chain perspective. *Healthcare, 10*(3), 479. https://doi.org/10.3390/healthcare10030479

Caspar, F. (2004). Technological developments and applications in clinical psychology and psychotherapy: Introduction. *Journal of Clinical Psychology, 60*(3), 221–238. https://doi.org/10.1002/jclp.10260

Castelnuovo, G., Gaggioli, A., Mantovani, F., & Riva, G. (2003). New and old tools in psychotherapy: The use of technology for the integration of the traditional clinical treatments. *Psychotherapy: Theory, Research, Practice, Training, 40*(1–2), 1–12. https://doi.org/10.1037/0033-3204.40.1-2.33

Cavanagh, K., Zack, J. S., Shapiro, D. A., & Wright, J. H. (2003). Computer programs for psychotherapy. In S. Gross, & K. Anthony (Eds.), *Technology in counselling and psychotherapy* (pp. 143–164). Macmillan Education UK.

Chaudhry, B. M. (2016). Daylio: Mood-quantification for a less stressful you. *mHealth, 2*, 34. https://doi.org/10.21037/mhealth.2016.08.04

Clough, B. A., & Casey, L. M. (2015). The smart therapist: A look to the future of smartphones and mHealth technologies in psychotherapy. *Professional Psychology: Research and Practice, 46*(3), 147–153. https://doi.org/10.1037/pro0000011

Collins, W. (2022). Mental health care. In W. Collins (Ed.), *Collins English Dictionary*. HarperCollins Publishers. https://www.collinsdictionary.com/us/dictionary/english/mental-health-care.

Conley, C. S., Raposa, E. B., Bartolotta, K., Broner, S. E., Hareli, M., Forbes, N., … Assink, M. (2022). The impact of mobile technology-delivered interventions on youth well-being: Systematic review and 3-level meta-analysis. *JMIR Mental Health, 9*(7), Article e34254. https://doi.org/10.2196/34254

Costescu, C. A., Vanderborght, B., & David, D. O. (2014). The effects of robot-enhanced psychotherapy: A meta-analysis. *Review of General Psychology, 18*(2), 127–136. https://doi.org/10.1037/gpr0000007

Darcy, A., Beaudette, A., Chiauzzi, E., Daniels, J., Goodwin, K., Mariano, T. Y., … Robinson, A. (2022). Anatomy of a Woebot®(WB001): Agent guided CBT for women with postpartum depression. *Expert Review of Medical Devices, 19*(4), 287–301. https://doi.org/10.1080/17434440.2022.2075726

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Formative Research, 5*(5), Article e27868. https://doi.org/10.2196/27868

November 1 DeAngelis, T. (2021). *Depression and anxiety escalate during COVID*. American Psychological Association https://www.apa.org/monitor/2021/11/numbers-depression-anxiety.

March 15 Delgone, M. C. (2022). *Woebot health secures $9.5 million investment from Leaps by Bayer. Business Wire* https://www.businesswire.com/news/home/20220315005370/en/Woebot-Health-Secures-9.5-Million-Investment-From-Leaps-by-Bayer.

Diano, F., Sica, L. S., & Ponticorvo, M. (2023). A systematic review of mobile apps as an adjunct to psychological interventions for emotion dysregulation. *International Journal of Environmental Research and Public Health, 20*(2), 1431. https://doi.org/10.3390/ijerph20021431

July 8 Dinneen, J. (2020). *I chatted with a therapy bot to ease my covid fears*. It was bizarre. *OneZero*. https://onezero.medium.com/i-chatted-with-a-therapy-bot-to-ease-my-covid-fears-it-was-bizarre-ccd908264660#:~:text=When%20Woebot%20launched%20in%202017,paid%20for%20as%20a%20therapeutic.

Donker, T., Griffiths, K. M., Cuijpers, P., & Christensen, H. (2009). Psychoeducation for depression, anxiety and psychological distress: A meta-analysis. *BMC Medicine, 7*(1), 79. https://doi.org/10.1186/1741-7015-7-79

Drum, K. B., & Littleton, H. L. (2014). Therapeutic boundaries in telepsychology: Unique issues and best practice recommendations. *Professional Psychology: Research and Practice, 45*(5), 309–315. https://doi.org/10.1037/a0036127

Economides, M., Ranta, K., Nazander, A., Hilgert, O., Goldin, P. R., Raevuori, A., et al. (2019). Long-term outcomes of a therapist-supported, smartphone-based intervention for elevated symptoms of depression and anxiety: Quasiexperimental, pre-postintervention study. *JMIR mHealth and uHealth, 7*(8), Article e14284. https://doi.org/10.2196/14284

Essig, T. (2015a). The gains and losses of screen relations: A clinical approach to simulation entrapment and simulation avoidance in a case of excessive internet pornography use. *Contemporary Psychoanalysis, 51*(4), 680–703. https://doi.org/10.1080/00107530.2015.1023669

December 14 Essig, T. (2015b). *Talkspace tarnishes promise of telehealth with extravagant claims*. Forbes https://www.forbes.com/sites/toddessig/2015/12/14/talkspace-tarnishes-promise-of-telehealth-with-extravagant-claims/.

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), Article e13216. https://doi.org/10.2196/13216

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), e19. https://doi.org/10.2196/mental.7785

Garfan, S., Alamoodi, A. H., Zaidan, B. B., Al-Zobbi, M., Hamid, R. A., Alwan, J. K., … Momani, F. (2021). Telehealth utilization during the Covid-19 pandemic: A systematic review. *Computers in Biology and Medicine, 138*, Article 104878. https://doi.org/10.1016/j.compbiomed.2021.104878

Gratzer, D., Strudwick, G., & Yeung, A. (2019). Mental illness: Is there an app for that? *Families, Systems & Health, 37*(4), 336–339. https://doi.org/10.1037/fsh0000451

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., et al. (2022). Mental health chatbot for young adults with depressive symptoms during the COVID-19 Pandemic: Single-blind, three-arm randomized controlled trial. *Journal of Medical Internet Research, 24*(11), Article e40719. https://doi.org/10.2196/40719

Isaacs-Russell, G. (2015). *Screen relations: The limits of computer-mediated psychoanalysis and psychotherapy*. London: Karnac Books Ltd.

October Jeong, S., & Breazeal, C. L. (2016). Improving smartphone users' affect and wellbeing with personalized positive psychology interventions. In *Proceedings of the Fourth International Conference on human agent interaction* (pp. 131–137). https://doi.org/10.1145/2974804.2974831.

Koh, J., Tng, G. Y., & Hartanto, A. (2022). Potential and pitfalls of mobile mental health apps in traditional treatment: An umbrella review. *Journal of Personalized Medicine, 12*(9), 1376. https://doi.org/10.3390/jpm12091376

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights, 11*, 1–9. https://doi.org/10.1177/1178222619829083

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lattie, E. G., Stiles-Shields, C., & Graham, A. K. (2022). An overview of and recommendations for more accessible digital mental health services. *Nature Reviews Psychology, 1*(2), 87–100. https://doi.org/10.1038/s44159-021-00003-1

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials. *World Psychiatry, 18*(3), 325–336. https://doi.org/10.1002/wps.20673

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. https://doi.org/10.1037/ccp0000459

Lipschitz, J. M., Van Boxtel, R., Torous, J., Firth, J., Lebovitz, J. G., Burdick, K. E., et al. (2022). Digital mental health interventions for depression: Scoping review of user engagement. *Journal of Medical Internet Research, 24*(10), Article e39204. https://doi.org/10.2196/39204

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. https://doi.org/10.1016/j.chb.2014.04.043

Mahtta, D., Daher, M., Lee, M. T., Sayani, S., Shishehbor, M., & Virani, S. S. (2021). Promise and perils of telehealth in the current era. *Current Cardiology Reports, 23*(9), 1–6. https://doi.org/10.1007/s11886-021-01544-w

Mello, F. L. D., & Souza, S. A. D. (2019). Psychotherapy and artificial intelligence: A proposal for alignment. *Frontiers in Psychology, 10*, 263. https://doi.org/10.3389/fpsyg.2019.002

Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Arean, P., Hollon, S. D., … Kaplan, R. (2009). The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychotherapy and Psychosomatics, 78*(5), 275–284. https://doi.org/10.1159/000228248

Mordoch, E., Osterreicher, A., Guse, L., Roger, K., & Thompson, G. (2013). Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas, 74*(1), 14–20. https://doi.org/10.1016/j.maturitas.2012.10.015

National Institute of Mental Health. (2017). *Depression in college students*. NIH Publication No. 15-4266. Retrieved from.

Nicol, G., Wang, R., Graham, S., Dodd, S., & Garbutt, J. (2022). Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: Feasibility and acceptability study. *JMIR Formative Research, 6*(11), Article e40242. https://doi.org/10.2196/40242

Norcross, J. C., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of Clinical Psychology, 67*(2), 127–132. https://doi.org/10.1002/jclp.20764

Oduntan, A., Oyebode, O., Beltran, A. H., Fowles, J., Steeves, D., & Orji, R. (2022). I let depression and anxiety drown me: Identifying factors associated with resilience based on journaling using machine learning and thematic analysis. *IEEE Journal of Biomedical and Health Informatics, 26*(7), 3397–3408. https://doi.org/10.1109/JBHI.2022.3149862

Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly, 93*, 249–253. https://doi.org/10.1007/s11126-022-09973-8

Prochaska, J. J., Vogel, E. A., Chieng, A., Baiocchi, M., Maglalang, D. D., Pajarito, S., … Robinson, A. (2021). A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug and Alcohol Dependence, 227*, Article 108986. https://doi.org/10.1016/j.drugalcdep.2021.108986

Ramachandran, M., Suharwardy, S., Leonard, S. A., Gunaseelan, A., Robinson, A., Darcy, A., … Judy, A. (2020). 74: Acceptability of postnatal mood management through a smartphone-based automated conversational agent. *American Journal of Obstetrics and Gynecology, 222*(1), S62. https://doi.org/10.1016/j.ajog.2019.11.090

Rathbone, A. L., Clarry, L., & Prescott, J. (2017). Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: Systematic review. *Journal of Medical Internet Research, 19*(11), Article e8598. https://doi.org/10.2196/jmir.8598

Riper, H., van Straten, A., Keuken, M., Smit, F., Schippers, G., & Cuijpers, P. (2009). Curbing problem drinking with personalized-feedback interventions: A meta-analysis. *American Journal of Preventive Medicine, 36*(3), 247–255. https://doi.org/10.1016/j.amepre.2008.10.016

Rizvi, S. L., Dimeff, L. A., Skutch, J., Carroll, D., & Linehan, M. M. (2011). A pilot study of the DBT coach: An interactive mobile phone application for individuals with

borderline personality disorder and substance use disorder. *Behavior Therapy, 42*(4), 589–600. https://doi.org/10.1016/j.beth.2011.01.003

Roesler, C. (2017). Tele-analysis: The use of media technology in psychotherapy and its impact on the therapeutic relationship. *Journal of Analytical Psychology, 62*(3), 372–394.

Shibata, T., & Wada, K. (2011). Robot therapy: A new approach for mental healthcare of the elderly–a mini-review. *Gerontology, 57*(4), 378–386. https://doi.org/10.1111/1468-5922.12317

Simmons, K., Garcia, E., Howell, M. K., & Leong, S. (2016). Personalizing, delivering and monitoring behavioral health interventions: An annotated bibliography of the best available apps. *Research Gate*. https://www.researchgate.net/profile/Sharlene-Jeffers/publication/309679218_Personalizing_Delivering_and_Monitoring_Behavioral_Health_Interventions_An_Annotated_Bibliography_of_the_Best_Available_Apps/links/581ce4e908ae12715af2136f/Personalizing-Delivering-and-Monitoring-Behavioral-Health-Interventions-An-Annotated-Bibliography-of-the-Best-Available-Apps.pdf.

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. https://doi.org/10.1001/archinte.166.10.1092

Thompson, C. (2018). *Joseph Weizenbaum, interviewed by clive Thompson for the New York times Magazine*. on April 18, 2002. Creative Commons https://creativecommons.org/licenses/by/2.0/.

Wachtel, P. L. (2010). Beyond "ESTs": Problematic assumptions in the pursuit of evidence-based practice. *Psychoanalytic Psychology, 27*(3), 251–272. https://doi.org/10.1037/a0020532

Wang, L., Mujib, M. I., Williams, J., Demiris, G., & Huh-Yoo, J. (2021). *An evaluation of generative pre-Training model-based therapy chatbot for Caregivers*. https://doi.org/10.48550/arXiv.2107.13115. *arXiv preprint arXiv:2107.13115*.

Watson, D., Clark, L. A., & Tellegan, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.

Wilson, E. A. (2010). *Affect and artificial intelligence*. University of Washington Press.

Zeavin, H. (2022). Therapy with a human face. *Dissent, 69*(1), 11–15. https://doi.org/10.1353/dss.2022.0002

Zilberstein, K. (2015). Technology, relationships, and culture: Clinical and theoretical implications. *Clinical Social Work Journal, 43*(2), 151–158. https://doi.org/10.1007/s10615-013-0461-2