Variable-dose support in an online mental health intervention:
A randomized, controlled exploratory study
by
Jack Brooks

Department of Psychology and Neuroscience
Duke University

Date: _____
Approved:

_____
Gary Bennett, Supervisor

_____
Rick Hoyle

_____
Gregory Samanez Larkin

_____
Dori Steinberg

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the Department of Psychology and Neuroscience
in the Graduate School of Duke University

2022

ABSTRACT

Variable-dose support in an online mental health intervention:
A randomized, controlled exploratory study
by
Jack Brooks

Department of Psychology and Neuroscience
Duke University

Date:_____
Approved:

_____
Gary Bennett, Supervisor

_____
Rick Hoyle

_____
Gregory Samanez Larkin

_____
Dori Steinberg

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Psychology and Neuroscience in the Graduate School of
Duke University

2022

Abstract

Digital health interventions are widely considered a highly promising solution to issues with access to evidence-based care. However, digital health interventions are limited by lower rates of engagement than are typically seen in traditional, face-to-face interventions. Despite the importance of engagement for intervention efficacy, engagement is rarely a primary outcome in intervention studies, and few studies have empirically tested intervention changes to improve engagement.

Hazel was a two-arm, randomized controlled exploratory study designed to investigate the impact of intervention design on engagement as a primary outcome. We conducted a trial of a mental health intervention, based on the Unified Protocol, delivered entirely online. Participants were randomized to complete the 12-week intervention self-guided (the unsupported arm) or with 4 weeks of therapist support (the supported arm). We sought to measure engagement as comprehensively as possible and therefore collected numerous self-report, behavioral, and objective measures of multiple facets of engagement, including how participants felt about the intervention, how they used the intervention technology, and how their behavior changed over the course of the intervention. We collected engagement outcomes at baseline, weekly during the intervention, and immediately following completion of the intervention. We hypothesized that a low dose of therapist support would improve engagement and subsequent mental health outcomes while being more scalable and feasible to implement than offering therapist support during all intervention weeks. The aims of this small, exploratory study were to learn more about patterns of engagement with the intervention and identify issues

with the current intervention and trial design. We used descriptive statistics and visualizations to understand and describe trends in the data rather than p-value significance testing, in line with best practices for pilot studies.

We recruited 23 North Carolina-based adults with clinically elevated depression and/or anxiety symptoms. Overall, the intervention and trial design appeared acceptable to participants. Participants in both arms had relatively positive attitudes toward the intervention at all time points, and there was no evidence that the response burden of our outcome measures was too high. However, several of our measures showed little variation between participants or over time. Our findings provides guidance for future studies to select alternative measures and/or to administer some engagement measures less frequently. This represents a meaningful step forward for the field of engagement research, as prior to this study, there was almost no empirical guidance on how to select measures of engagement or how frequently to administer them.

We observed a substantial drop in engagement for participants in the supported arm when video sessions began, contrary to our expectations that video sessions would improve engagement. Notably, video sessions began partway through the intervention and participants were blinded to condition until the onset of these sessions; changes to blinding and/or timing of video sessions might improve engagement. At the same time, those participants in the supported arm who did engage in video sessions (n=4) completed more of each weekly lesson than participants in the unsupported arm, indicating that there is evidence that therapist support can increase engagement.

Secondary outcomes were sample diversity and depression and anxiety symptom change. Our sample was comparably diverse to other digital health studies in terms of race and socioeconomic status, and we overrecruited LGBTQ+-identified participants. Despite finding little evidence of changes in behavior or the theoretical mechanisms underlying the intervention, the majority of participants showed improvement in their symptoms.

Contents

List of Tables

List of Figures

Acknowledgements

In her essay "Allegiance to Gratitude," Dr. Robin Wall Kimmerer quotes the extensive Thanksgiving Address of the Onondaga Nation, thanking in turn the earth, the water, the fish, the plants, the berries, the Three Sisters, the medicines, the trees, the animals, the birds, the four winds, the thunder and lightning, the sun, the moon, the stars, the teachers, the spirits, and everything that has been unintentionally left out. Dr. Kimmerer notes that sometimes listeners appear impatient with how long the address can go on for. "'Poor you,'" she writes to these individuals. "'What a pity that we have so much to be thankful for.'" Writing this, I am feeling exceptionally "poor me."

First, thank you to Dustin, without whom none of this would have been possible. I simply do not have the words, which will say a lot to anyone who has ever met me. Next, thank you to Hina and Shayla, who have made my family and my heart somewhere between three times and infinitely as big.

Thank you Gary, for making the past 6 years possible. I honestly cannot imagine going through this process anywhere else, with anyone else. Thank you for talking me out of dropping out and going to law school (repeatedly), for offering to help me figure out how to get someone to pay me to go to law school (once), for allowing me the freedom and the flexibility to pursue the things I am truly interested in (always). It has truly, indelibly, shaped who I am as a person. Relatedly, all complaints about who I am as a person can be routed to gary.bennett@duke.edu in perpetuity, thanks.

Thank you Rick, Greg, and Dori for the extensive support and encouragement you have provided me over the years. Thank you also for the gentle advice to scale back when I have set unreasonable goals, and then gently advising me again to scale back when I did not listen the first time (and again, and again…).

xiv

Thank you to everything and everyone that has been unintentionally left out. It is a blessing to have such abundance you can take even a single ounce for granted. This project was financially supported by a Charles Lafitte Foundation Graduate Grant Award for Psychological Research related to Identity, Diversity, Inclusion, Equity, and Thriving and the Stella Powell-Williams Grant for Research Impacting Community.

Chapter 1. Background

In recent years, healthcare interventions leveraging the widespread availability of sophisticated technology such as smartphone applications, websites and text messaging have risen as a promising potential solution to the inaccessibility of quality healthcare.[1] As of January 2018, 77% of US residents own a smartphone and another 18% own other mobile phones.[2,3] Globally, 59% of the world's population own smartphones and another 31% own other mobile phones.[4,5] Rates of access are even higher when computers, tablets, wearable devices and community sources such as libraries, schools and internet cafes, are considered. This field, collectively referred to as digital health, has been embraced by researchers, healthcare providers and patients alike for the treatment of a variety of physical and mental health conditions.[6] Research has found that digital health interventions can be highly effective for a variety of physical and mental health issues.[7–14]

Digital health interventions are particularly promising to resolve health disparities for people who do not have access to traditional in-person care. However, digital interventions (as well as in-person interventions) tend to strongly oversample white, highly educated, and high-income participants; findings in these populations may not be replicable in racially/ethnically diverse or low-SES populations. Further, few digital health studies report on the inclusion of sexual and gender minority participants, but it is well-established that these populations also struggle to access affirming care in in-person settings.[15] In recent studies, 30.6% of LGBTQ+ participants reported accessing care is somewhat to very difficult; 68% of transmasculine adults and 23.98-46.66% of transfeminine adults (depending on setting) report experiences of mistreatment in healthcare settings; and 100% of non-binary and gender non-conforming individuals in a

2018 study reported difficulties accessing affirming care even in specific clinics for transgender patients.[16–18] Digital health has the unique potential to connect people from marginalized communities with unbiased, affirming care no matter where they live, but this potential has thus far been underexplored.

This promising field remains further limited by much higher rates of treatment drop-out than are seen in traditional in-person interventions. Among all publicly available health apps, recent research indicates that only 31% of users are still active the first day after downloading; after a month, only about 15% of downloaders are still using the app.[19] In digital health research, intervention usage and non-usage is often referred to as "engagement."

1.1 On "Engagement"

Efforts to interpret and understand the role of engagement in digital health are complicated by the multiple ways that this term has been used in the literature. Notably, low engagement is considered an issue in digital health across these disparate definitions. Among the most-cited reviews of digital health engagement, "engagement" is variously defined as: a preliminary intervention stage where participants' readiness to change is assessed and they are made aware of and recruited to the intervention; an individual state of active involvement in one's own health and health management; the ways in which an individual uses and interacts with a digital health intervention; the progress an individual makes toward achieving behavior-change goals; and a subjective experience of focus and attention while interacting with an intervention.[20–25] The findings regarding engagement are further complicated by the overlap between these definitions (e.g., engagement as enjoyment) and potential outcomes (e.g., engagement

increases enjoyment). Several of these reviews note that engagement is multiply defined, and two attempt to reconcile the considerable diversity of the field in their definitions.

Yardley et al (2016) identified two prominent definitions of engagement across individual studies — engagement as the process of interacting with the intervention and engagement as the process of achieving behavior-change goals — and conceptualized these as micro and macro levels of engagement, respectively.[24] In this model, engagement starts on the micro level, with interactions with the digital intervention in the absence of behavior change. With sufficient micro-engagement, the individual will begin to exhibit signs of behavior change, or macro level engagement, though this behavior change is still dependent on micro-level engagement with the intervention. Eventually, behavior change becomes normalized as part of the person's daily routine and engagement with the intervention is no longer necessary to maintain behavior change; however, if macro-engagement begins to falter it may be necessary to return to the digital intervention and micro-engagement to prevent or treat relapse in behavior. The tipping point at which engagement with the intervention ceases to be necessary for health behavior change highlights the limitations of defining engagement solely on the micro level, as there is no way to distinguish between individuals who cease engaging with an intervention without achieving behavioral change (e.g., intervention drop-out and non-use attrition) from those who cease engaging *because* of behavior change (i.e., intervention success).

Similarly, Perski et al (2016) identified two common definitions of engagement across computer science, human-computer interaction (HCI) and behavioral science literatures: engagement as a subjective experience marked by interest, enjoyment and a sense of flow, and engagement as the usage behaviors an individual shows toward the

digital intervention.[25] The authors theorize that the subjective user experience of the

intervention is, at least in part, responsible for the observable usage behaviors, and they

also note that engagement may vary both between and within individuals over time.

Perski et al's definition of engagement as intervention usage behaviors is notably similar

to Yardley et al.'s definition of micro-level engagement, resulting in three leading

definitions of engagement:

1. *Experiential engagement*: the subjective experience of using an intervention;

2. *Technological engagement:* the specific usage of intervention hardware/software; and

3. *Behavioral engagement:* the resultant process of behavior change toward health goals.

These facets of engagement are not discrete and independent, but rather are interrelated

processes that dynamically impact one another and subsequent health outcomes.[24,25] The

importance of behavior change in improving physical and mental health outcomes is

well-established; it is the foundation of psychology and behavioral medicine

interventions.[26,27] The impact of technological engagement on behavior change (and

subsequently on health outcomes) has similarly been established in the extant

literature.[24,25] Some level of exposure to the intervention is necessary to develop and

maintain behavior change, though the exact dose necessary for behavior change or

improved outcomes is unclear. Additionally, the point at which technology usage is no

longer necessary to maintain behavior change is unknown.

We theorize that there are complex, reciprocal relations between these facets of

engagement and ultimate health outcomes (see **Figure 1**). We conceptualize experiential

engagement as the most "upstream" facet of engagement, with participants' perspectives

and attitudes toward the intervention being formed from the first moments of use. We theorize that participants who find an intervention more interesting, enjoyable, and immersive will show greater technological engagement, and that greater technological engagement will spill over into greater behavior change and subsequently greater improvements in health outcomes. At the same time, we expect that with greater technological engagement may improve experiential engagement through increased mastery of the intervention platform. We also theorize that greater behavior change and greater symptom change will improve experiential engagement, due to more positive outcome expectancies about the intervention as participants see target behaviors increasing and symptoms decreasing. Similarly, we expect that greater symptom change will reinforce behavior change and improve behavioral engagement. Finally, we expect greater behavior change and symptom change to have an inhibitory effect on technological engagement, as intervention usage no longer becomes necessary to maintain treatment progress.
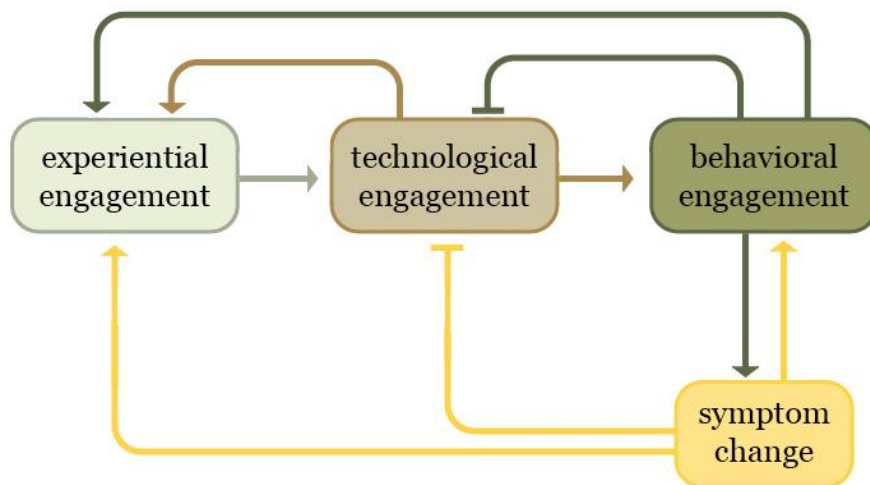
Figure 1: Theoretical relationship between facets of engagement and health outcomes. Triangular arrowheads represent positive impacts; flat arrowheads represent inhibitory relations.

1.2 Predictors of Engagement

Studies have found that these three facets of engagement are associated with a number of participant factors (e.g., motivation to change, accurate expectations of the intervention, computer literacy), intervention factors (e.g., level of user control, novelty, complexity) and participant-intervention match factors (e.g., cultural norms, financial and material resource demands, time commitment).[24,25] Intervention and participant-intervention match factors are particularly promising for promoting greater engagement, as these factors are more easily manipulated by intervention developers.

However, attempts to understand the impact of intervention design on any facet of engagement are limited by the high variability in how interventions are described — an issue affecting the field of behavioral medicine broadly. For this reason, Michie et al. (2013) developed the Behavior Change Techniques (BCT) taxonomy, an evidence-based list of 93 intervention components associated with behavior change.[28,29] It was designed to improve reliability in intervention descriptions across researchers and fields of research, in accordance with CONSORT guidelines for precise reporting of intervention components. The taxonomy was developed by an international panel of behavior change experts who evaluated the labels and definitions of every known intervention component taxonomy, grouped them based on similar mechanisms of change, and performed reliability testing by applying the resulting taxonomy to published intervention descriptions. The finished taxonomy can be used at the intervention design stage or to code published studies for reviews or meta-analysis.

By definition, all BCTs can impact behavioral engagement. In a recent review of the impact of BCTs on experiential engagement, Brooks et al. (unpublished) found the strongest support for 5 BCTs: self-monitoring outcomes of behavior (6 studies available),

feedback on outcomes of behavior (5 studies), framing/reframing (5 studies), self-monitoring of behavior (4 studies), and behavioral practice/rehearsal (4 studies), as all available studies found greater experiential engagement when these BCTs were included.[15] The more trials finding favorable effects of a BCT, the more probable it is that the BCT's impact on engagement is resilient across different constellations of included BCTs and treatment populations. However, experiential engagement was the primary outcome for only two of the 43 studies included in this review, meaning than the vast majority of included studies were not powered to find a significant effect on engagement if there was one. While we can never be fully certain that non-significant results reflect a true lack of difference, we are especially limited in interpreting non-significant results for studies that lack power to find significance.

Overall, the results of this review suggest that adding *any* BCT generally improves experiential engagement, and relatively few BCTs were harmful to experiential engagement. In most cases where the presence of a BCT was tied to lower engagement, it was relative to other — and often more — BCTs in the arm with higher engagement. For instance, there were few cases in which a given BCT was the only difference between arms *and* the arm with that BCT had poorer engagement. Rather, it was more often the case that one arm containing a small number of BCTs showed worse engagement than an arm with substantially more BCTs, making it inaccurate to conclude the BCTs in the less-engaging arm *caused* the lower rate of engagement. This is consistent with "black box" effects observed in other intervention outcomes, in which simply adding more components can generally lead to greater observed change. Finally, it is possible that BCTs would have positive impacts on some measures of experiential engagement and not others, making it difficult to draw conclusions across studies that use different outcomes.

However, there are several important limits to this study. As noted above, only two of the 43 included studies were powered to find a difference in experiential engagement. Further, none of the included studies were designed to empirically evaluate or compare BCTs. Instead, the included studies were designed to compare different treatment modalities (e.g., cognitive versus behavioral treatment for depression), more and less intensive versions of the same treatment (e.g., treatments with and without support), or small differences in the implementation of a treatment (e.g., using one's own phone versus a study-provided phone for the same mobile intervention). These differences often led to corresponding differences in BCTs, which we coded and used for our analyses; however, because this was not the investigators' intention there is no guarantee that factors other than these BCT differences were held constant between arms. Similarly, most of our eligible studies added multiple BCTs to an arm simultaneously, making it impossible to attribute observed differences to any individual BCT or to interaction of multiple BCTs, substantially limiting the interpretability and generalizability of these results.

Finally, though not all studies reported on participant demographics, the majority of participants were white, female, employed full-time and had at least a college degree. The relative non-representativeness across these studies is noteworthy, as digital health is often considered especially promising for populations with limited access to traditional in-person care. Due to these significant limitations, the findings of this study are best used for hypothesis generation, rather than drawing any conclusions about the impact of individual BCTs on engagement or providing recommendations for intervention design.

1.3 Social Support

Social support is one particularly well-established predictor of all facets of engagement, as well as ultimate health outcomes. As noted above, Brooks et. al.'s review found preliminary evidence suggesting that social support may positively impact experiential engagement in at least some scenarios. In the BCT framework, social support includes "non-contingent praise or reward," encouragement, and counseling aimed at promoting the desired behavior. There are three subtypes of social support: emotional support, practical support, and unspecified support (used when there is insufficient information to determine whether support was emotional, practical, or both). Below, we summarize results from the review related to all three types of social support.

Social support was the most-tested BCTs in this review. There were a total of 14 trials that included any of the three types of social support; 12 of these 14 could best be classified as unspecified social support. Of these 14 trials, seven found greater experiential engagement in the arms that included unspecified social support. Five found minimal differences between arms, and two found mixed results. Notably, for four of the 14 studies, social support was the only difference between arms. For these studies, all differences between arms can be attributed to the presence or absence of support, making them particularly informative. See **Table 1** for a summary of study results.

These mixed results may be driven by other, more impactful BCTs included in these trials; they may reflect interaction effects with other included BCTs or simply the presence of more BCTs in certain arms; they may be due to different implementations of the BCT between trials; or they may be due to non BCT-related differences between arms: other intervention design features not held constant across arms, intervention

delivery methods, intervention effectiveness, the health condition or behavior being targeted, factors related to the included population, or the measure of experiential engagement used.

For one study, the inclusion of social support was the only difference between arms and produced significantly greater experiential engagement. Tonkin-Crine et al. (2013) compared arms with and without social support in a CBT-based intervention for irritable bowel syndrome.[50] They found that participants in the arm without social support were more likely to describe limited engagement with all components of the intervention, while participants in the supported arm were more likely to describe engaging with the practical advice and/or psychological information. An additional six studies with multiple BCT differences between arms found results in favor of unspecified social support. Beiwinkel et al. (2017) compared an information control arm with an active intervention arm, which included social support, for the treatment of depression and workplace absence, and found that participants were significantly more satisfied with the intervention in the arm that included social support.[30] Berger et al. (2017) found significantly greater experiential engagement in the arms of their social phobia intervention that included social support.[31] Fitzpatrick et al. (2017) found greater experiential engagement in the arm of their depression intervention that included unspecified social support for treatment of depression.[32] Johnston et al. (2016) found significantly greater experiential engagement in the arm of their behavior change intervention that included social support.[33] Morris et al. (2015) found significantly greater experiential engagement in the arm of their depression intervention that included social support.[34] In a smoking cessation intervention, Mcclure et al. (2016) found greater experiential engagement in the arm that included social support.[35]

On the other hand, two studies for which unspecified social support was the only difference between arms found no major differences in experiential engagement between arms. Ho et al. (2014) compared CBT treatments for insomnia with and without support.[36] They found minimal differences between arms in treatment acceptability or satisfaction between arms at any time point. Similarly, in an intervention to increase physical activity levels, Alley et al. (2016) compared arms with and without support and found minimal differences between arms.[37] Three additional studies found minimal differences between arms, though these included multiple BCT differences between arms, making it less clear how attributable their results are to social support versus other differences. In a depression treatment trial, Mira et al. (2017) compared an arm with automated support only to an arm with both automated and human support.[38] Both before and after treatment, there were minimal differences between arms on perceived treatment logic, satisfaction, willingness to recommend, personal usefulness or usefulness for other problems. Similarly, Unick et al. (2016) compared weight loss interventions with and without practical support, and found minimal differences in program satisfaction between arms.[39] Hales et al. (2016) compared a weight loss mobile application with and without emotional support; there were no significant differences in intervention satisfaction between arms.[40]

In comparison, two studies found mixed results. For Sundstrom et al. (2016), social support was the only difference between arms. They compared three self-help arms for the treatment of problematic alcohol abuse: an arm without social support; an arm with asynchronous text-based support; and an arm with where participants could choose between synchronous and asynchronous text-based support.[41] They found that participants in the two supported arms were more likely to say the intervention was

personal and effective, and they were more willing to recommend the intervention than participants in the unsupported arm. Participants in all arms reported similar usefulness for increasing insights into risks of drinking. Finally, Watson et al. (2012) found that more participants in the arm that included social support perceived the intervention as beneficial for changing exercise and dietary habits in a physical activity intervention, with no difference between arms on overall perceived benefit from participating in the intervention.[42]

Table 1. Summary of findings on the impact of social support on experiential engagement.

Study	Type of Support	Only Difference Between Arms?	Result
Tonkin-Crine 2013	Unspecified	Yes	Increased engagement
Beiwinkel 2017	Unspecified	No	Increased engagement
Berger 2017	Unspecified	No	Increased engagement
Fitzpatrick 2017	Unspecified	No	Increased engagement
Johnston 2016	Unspecified	No	Increased engagement
Morris 2015	Unspecified	No	Increased engagement
Mcclure 2016	Unspecified	No	Increased engagement
Ho 2014	Unspecified	Yes	Minimal differences
Alley 2016	Unspecified	Yes	Minimal differences
Mira 2017	Unspecified	No	Minimal differences
Unick 2016	Practical	No	Minimal differences
Hales 2016	Emotional	No	Minimal differences

| Sundstrom 2016 | Unspecified | Yes | Mixed |
| Watson et al. (2012) | Unspecified | No | Mixed |

The considerable variation in the findings of these studies suggests that the impact of unspecified social support may be moderated by the population or the other BCTs that are included in the intervention. There is also mounting evidence that interventions with social support may increase rates of technological engagement, and a substantial body of research indicates that digital health interventions with social support produce greater changes in health behaviors and outcomes than unsupported or "standalone" interventions.[15,43–45]

Notably, almost all of this research has compared arms with no support to arms with frequent, high doses of support throughout the entire intervention period. However, there are substantial dissemination issues when including continuous support in digital interventions. Many digital health interventions developed by researchers include a human counseling component; however, these interventions are rarely made available to the public after testing and there are significant limitations to doing so as typically designed. Conversely, most publicly available digital health treatments are standalone interventions created by tech companies with limited empirical evidence for their efficacy, often including intervention elements that are actively contraindicated by best practices for the condition being targeted.[44,46,47]

1.4 Present Study

In an effort to find a compromise between the competing issues of scalability and efficacy, this study sought to compare the impact of a low dose support versus no support

on engagement and health outcomes in a digital health intervention. Further, we collected measures of all three facets of engagement as primary outcomes. Our primary aims for this study were to begin exploring the impact of our intervention on engagement and health outcomes and to determine the feasibility and acceptability of the trial design in a small sample.

1.5 Theoretical Constructs

The Model of Supportive Accountability theorizes that human inclusion improves outcomes through the emotional connection between the user and provider, the provider's ability to hold the user accountable, and the perceived legitimacy of the provider.[48] In this model, accountability can be further attributed to the social presence of another person, clear expectations for the user, collaborative and flexible goal-setting, and benevolent performance monitoring; legitimacy includes the user's sense that the provider is trustworthy, well-intentioned, and holds expertise, as well as clearly-defined roles for both the provider and the user. Our conceptual model (**Figure 2**) draws on the Model of Supportive Accountability to explain the relationship between the dose of human support, engagement, and symptom change. For the present study, we theorized that the presence of a low dose of support would affect all three facets of engagement, compared to no support. Specifically, we anticipated that the perceived legitimacy of the provider would have a positive effect on all three facets of engagement, while the bond between provider and participant would have a positive effect on experiential and technological engagement and the accountability of a provider would have a positive effect on technological and behavioral engagement.

model of supportive accountability
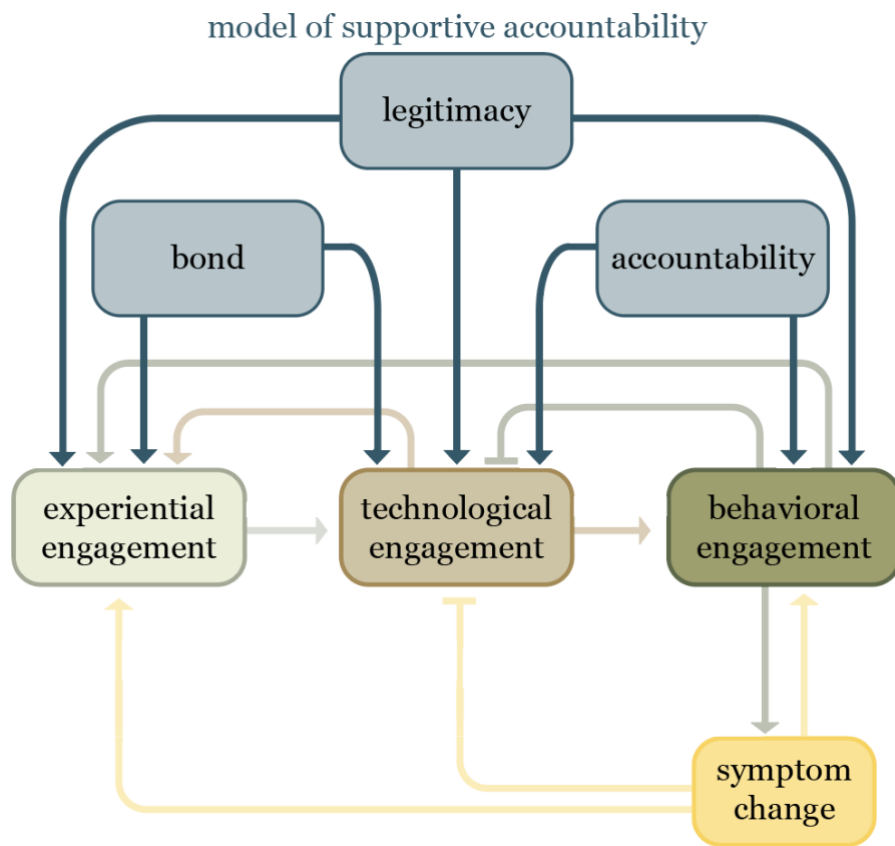
Figure 2: Conceptual Model.

Chapter 2. Methods

This exploratory trial aimed to investigate how a low dose of human support impacts engagement and health outcomes in a digital mental health intervention. We collected measures of experiential, technological and behavioral engagement as primary outcomes. Each of these facets of engagement was operationalized through multiple measures, including self-report and objective behavioral data, at multiple timepoints throughout the intervention. The goals of this study were 1) to assess the feasibility and acceptability of our intervention design and the feasibility of our trial methods in a small sample to inform the design of a larger, fully powered trial and 2) to begin to better understand and generate data-informed hypotheses about the impact of a low dose of support on engagement and health outcomes.

2.1 Design Overview

We conducted a 16-week, randomized controlled trial comparing an intervention for depression and anxiety symptoms with two different levels of support. Eligible participants were randomized to receive: 1) a 12-week, online mental health intervention ("Hazel") with a low dose of human support (the "supported" arm); or 2) the Hazel intervention without human support at any point (the "unsupported" arm). Data were collected at baseline, weekly during the intervention period, and at 1-month follow-up. This study was approved by the Duke University Campus Institutional Review Board (protocol number 2021-0292) and registered at ClinicalTrials.gov (NCT04810988).

The Hazel intervention was based on the Unified Protocol, an empirically supported, transdiagnostic treatment for emotion disorders.[49] We operationalized

human support according to the BCT taxonomy definition of social support, including "non-contingent praise or reward," encouragement, and counseling aimed at promoting the desired behavior. We chose to operationalize our low dose of human support as four weeks based on the specifics of the Hazel intervention, as described in greater detail below.

Our primary outcomes of interest were engagement with the intervention, operationalized in a number of different ways and measured repeatedly throughout the course of the intervention, in order to capture the complexity of "engagement" as a construct (see "Outcomes" for more details). Our secondary outcomes were depression and anxiety symptoms, measured at baseline, weekly during the intervention, and 1-month follow up.

2.2 Intervention Design

The Hazel intervention was based on the Unified Protocol, an empirically supported, transdiagnostic treatment for emotion disorders based on principles of cognitive behavioral therapy.[49] The Unified Protocol was initially developed and tested as a face-to-face, individual therapy which could be delivered by a trained therapist via weekly one-hour sessions over the course of 12 to 18 weeks. The Unified Protocol focuses on increasing awareness of maladaptive thoughts and behaviors while decreasing avoidance of emotions and physiological sensations. Each session starts with self-monitoring of depression and anxiety symptoms using a symptom survey, followed by provision of psychoeducation and in-session practice exercises related to that week's topic. At the end of each session, patients are assigned a home practice assignment to increase adaptive coping and generalize skill use outside of the therapy room. Their experiences of the home practice are reviewed at the start of the next session. The Unified

Protocol is broken down into 8 content modules, six of which are considered mandatory and two of which are considered optional.

Module 1 is focused on increasing motivation, readiness and self-efficacy for change. During this module, participants set short-, medium- and long-term goals and explored the pros and cons of change. Module 2 focuses on increasing understanding and awareness of emotions. Participants received psychoeducation about emotions and began self-monitoring their emotions. Module 3 is focused on further expanding awareness of emotions. Participants used multimedia sources to induce emotion and mindfully observed the interplay of thoughts, physical sensations, behavior and emotions. Module 4 focuses on the role of *thoughts* in emotional disorders. In the first week of this module, participants learned about common cognitive distortions and how those distortions can impact emotions, and then practiced identifying their own distorted thinking in the moment. In the second week, participants learned ways of challenging distorted thoughts and thinking more flexibly. Module 5 is focused on the role of *behaviors* in emotional disorders. Participants identified their maladaptive, emotion-driven behaviors in order to change them. Module 6 focuses on the role of physical sensations in emotional disorders. Participants completed interoceptive exposure exercises to observe the relation between physical sensations, thoughts and behavior. Module 7 focuses on exposure to interoceptive and situational triggers for emotions. Participants created an exposure hierarchy to identify specific triggers and then completed exposure exercises to increasingly difficult items from the hierarchy. Module 8 is focused on maintenance of treatment progress and relapse prevention.

The Hazel intervention included all eight modules of the Unified Protocol; there were no substantive changes to module content between the Unified Protocol and the Hazel intervention. All materials for the Hazel intervention were based on the Unified Protocol Workbook, a publicly available workbook including all psychoeducation, practice exercises, symptom self-monitoring surveys, and home practice assignments for the Unified Protocol.[50] This workbook was designed to be used either as a standalone, self-guided tool or in conjunction with a therapist. The Hazel intervention modified the Unified Protocol Workbook in two ways. First, we edited the vocabulary and sentence structure to be as close to a fifth grade reading level as possible, based on evidence that 52% of US residents read English at a 5th-grade level or below.[51] Second, we replaced the examples in the Unified Protocol Workbook with more diverse examples. We edited content to include names that are more commonly associated with BIPOC communities, to include a variety of gender identities and sexual orientations, and to include scenarios which would be more familiar to people from lower socioeconomic backgrounds (e.g., an example of a man who is worried about the cost of fixing a flat tire). All materials for the Hazel intervention were reviewed by one of the authors of the Unified Protocol manual to ensure fidelity to the Unified Protocol.

2.3 Experimental Manipulation

Participants were randomized to receive the Hazel intervention either with or without four weeks of social support. For participants in the supported arm, support was provided during the emotion exposure module (weeks 8-11). Participants were blinded to condition until Week 7, when those in the supported arm received an email introducing their assigned study therapist. This module involves systematically engaging in activities that produce significant distress, with the goal of learning to tolerate distress and decreasing avoidance of valued activities. We chose this time period to provide support,

rather than the start of the intervention, because the emotion exposure module is one of the most emotionally challenging and complex parts of the intervention. We hypothesized that unsupported attempts at complex tasks may present particularly high risks of negative attitudes, non-adherence, incorrect practice, or non-use attrition, and that the therapist support may therefore be particularly beneficial during such tasks. Similarly, we decided to operationalize a "low dose" of support as four out of 12 weeks to align with this four-week-long module.

2.3.1 Similar Components Across Treatment Arms

Participants in both arms received the 12-week Hazel intervention delivered via the same web platform (RedCap). Every week, participants received an emailed message with a link to that week's lesson. Starting in the second week, participants who had not completed the previous week's lesson received a reminder email including a link to that week's lesson. These emails were delivered by a depersonalized study email account. The weekly links directed participants to complete the following intervention components, in order:

1. depression symptom survey,
2. anxiety symptom survey,
3. psychoeducational reading,
4. practice exercises,
5. home practice instructions, and
6. home practice monitoring logs.

Each week's psychoeducational reading, practice exercises and homework assignments were based on the weekly topic, as described above. Following completion of each week's intervention content, participants were prompted to complete measures related to engagement outcomes. The psychoeducational reading, in-session practice exercises, and

home practice assignment for each week of the Hazel intervention are summarized in

Table 2, below.

Table 2. Content of the Hazel intervention, by week.

Mod.	Wk.	Psychoeducation	Practice Exercises	Homework
1	1	-Motivation & ambivalence about change -How to set effective goals	-Pros/cons of change -Treatment goals	
2	2	-Function of emotions -Three component model of emotions -Emotion-driven behavior & learned responses -The ARC of emotions	-Describe a time when negative emotions were helpful -Identify three components of current emotion -ARC of emotions example	-Three Component Model of Emotions worksheet -EDBs in Context worksheet
3	3	-Present-focused awareness -Nonjudgmental awareness of emotions -Primary versus secondary emotions (check the facts)	-mindfulness exercise -reaction to own emotions versus reaction to friend's emotions -emotion induction	-2x daily mindfulness exercises -EDBs in context worksheet -non-judgmental, present-focused awareness worksheet -Anchoring in the present worksheet -mood induction recording worksheet
4	4	-Automatic appraisals -thinking traps	-ambiguous picture exercise -downward arrow technique	-downward arrow technique -identifying automatic appraisals worksheet
	5	-reappraisal and cognitive flexibility -workability versus "rightness"	-generating alternative interpretations exercise	-identifying and evaluating automatic appraisals

5	6	-short & long-term consequences of emotion avoidance -types of avoidance -adaptive versus non- adaptive emotion- driven behaviors -countering maladaptive behaviors	-identify predictors of avoidance, short- and long- term consequences -thought suppression exercise -imaginal avoidance -identify maladaptive EDBs -generate alternatives to maladaptive EDBs	-list of emotion avoidance strategies -practice countering maladaptive EDBs
6	7	-avoidance of physical sensations	-identify avoided/distr essing physical sensations -induce symptoms & practice non-judgmental awareness	-symptom induction practice worksheet
7	8	-introduce & orient to emotion exposures	-develop exposure hierarchy	
	9		-emotion exposure & processing	-in vivo exposure
	10		-emotion exposure & processing	-in vivo exposure -design future exposures
	11		-emotion exposure & processing	-in vivo exposure -design future exposures
8	12	-relapse prevention and maintenance -warning signs -skill generalization	-review skills, triggers, treatment progress -goal setting	

2.3.2 Unsupported Arm

Participants randomized to the unsupported arm received the intervention exactly as described above, with no additional components. Completion of all weeks was entirely self-guided, to emulate a publicly available, standalone intervention.

2.3.3 Supported Arm

Participants randomized to the supported arm received all aspects of the intervention described above, with no differences between arms for the first seven weeks of the intervention (Modules 1-6). Participants in this arm were introduced to their assigned study therapist in week 8, at the start of Module 7, via email. During Module 7 (exposure; weeks 8-11), participants in this arm received synchronous, video-call contact with their assigned study therapist for live demonstration of exposure exercises, monitoring and feedback of participant exposure exercises, and post-exposure processing. These video sessions lasted approximately 50 minutes per participant per week. For Module 8 (relapse prevention; week 12) participants in the Supported arm completed the intervention independently, without any therapist support.

Table 3. Unique intervention components by module and arm.

Module	Unsupported	Supported
1		
2		
3		
4		
5		
6		

7	self-guided development of exposure hierarchy, exposure exercises and post-exposure processing	live video demonstration, monitoring, feedback and support of exposure exercises
8		

2.3.4 Study Therapists

Support was provided by trained study therapists. Study therapists were students enrolled in Duke's clinical psychology Ph.D. program. All therapists were trained in the study protocol, the Unified Protocol, and exposure therapy. Study therapists also attended weekly group supervision for the duration of the intervention. Trainings and supervision were led by the primary investigator (JB), with umbrella supervision provided by licensed clinical psychologists at Duke University. One of the study supervisors was a coauthor of the Unified Protocol manual.

2.4 Participants

We recruited participants meeting the inclusion and exclusion criteria listed in **Table 4**. To be eligible, individuals needed to be adults living in North Carolina, with access to the Internet and a computer, smartphone and/or tablet for the duration of the study. They needed to be fluent in reading and writing in English and have moderately severe or impairing symptoms of depression (operationalized as an Overall Depression Severity and Impairment Scale score of 8 or higher) and/or anxiety (operationalized as an Overall Anxiety Severity and Impairment Scale score of 8 or higher).

Individuals were excluded if they were experiencing current suicidality (operationalized as a positive result on the Ask Suicide-Screening Questions tool);

current psychosis (measured by the SCID-5 psychotic disorders subsection); current substance abuse disorders other than nicotine, cannabis or caffeine (measured by SCID-5 substance use disorders subsection); were clinically underweight (operationalized as a BMI of 18.5 or lower); or had engaged in self-harm behaviors in the past 12 months. Individuals who were receiving other psychotherapy or taking psychiatric medications, or who planned to start other therapy or psychiatric medications during the course of the intervention, were also excluded, as were any individuals who had received 8 or more sessions of any cognitive behavioral therapy in the past 12 months.

We sought to over-recruit participants of minoritized backgrounds who tend to be underrepresented in digital health research; we used targeted recruitment strategies to increase racial/ethnic diversity and LGBTQ+ inclusion in our sample. We further used inclusive language and examples on all measures and intervention materials, as described above.

2.5 Recruitment

We used a multi-pronged recruitment strategy, consisting of (1) social media; (2) community organizations; (3) professional organizations; and (4) snowball recruitment. We recruited throughout the state of North Carolina, as the intervention included no in-person elements and could be completed from anywhere. To assist in recruiting a diverse sample, we included both broadly used social media (Twitter, Instagram, and Reddit) and platforms that cater to LGBTQ+ populations (Scruff and Lex). We also shared recruitment materials with providers at community health clinics in the area and local organizations that serve primarily BIPOC and LGBTQ+ communities. Lastly, we posted recruitment materials with professional organizations such as the Duke Clinical Trials Registry, the National Clinical Trials Registry, and ResearchMatch. Advertisements

included a brief description of the study and eligibility criteria and a link to the study website, where more information was available. Participants were enrolled in the study on a rolling basis through our recruitment window (April 2021 through December 2021), with a plan to cap enrollment at a maximum of 50 participants for feasibility reasons.

Table 4. Inclusion/Exclusion Criteria.

Inclusion Criteria	
	18+ years old
	Live in North Carolina
	Fluent in reading and writing English
	Access to Internet/data plan and Internet-capable device throughout intervention
	Overall Depression Severity and Impairment Scale (ODSIS) **and/or** Overall Anxiety Severity and Impairment Scale (OASIS) summary score >= 8
Exclusion Criteria	
	Currently receiving other psychotherapy or planning to receive other psychotherapy during course of intervention
	>=8 sessions of CBT in past 12 months Currently taking psychiatric medications or planning to take psychiatric medications during course of intervention
	Current suicidality, operationalized as a positive result on the Ask Suicide-Screening Questions (ASQ) Tool
	Self-harm in past 12 months
	Current psychosis, as measured by SCID-5 psychotic disorders subsection
	Substance abuse (other than nicotine, cannabis or caffeine) in the past 12 months, as measured by SCID-5 substance use disorders subsection (past 12 months only)
	BMI <= 18.5

2.6 Enrollment Procedures

Online screening. The study website directed interested individuals to an online screener questionnaire to assess initial eligibility based on age, Internet and Internet-capable device access, treatment history, diagnostic contraindications, and availability over the next 4 months. Participants were also administered the Overall Depression Severity and Impairment Scale (ODSIS) and the Overall Anxiety Severity and Impairment Scale (OASIS) to confirm symptom eligibility.

Telephone screening and consent. Participants who were eligible based on the online screener were prompted to enter their contact information and then contacted by study personnel within 3 business days of completing the online screener. The telephone screening included an overview of study details, provision of additional information, and confirmation of individuals' interest and eligibility. Those who were eligible based on the telephone screener were emailed a link to the study consent form to complete during the screening call.

2.7 Randomization

Immediately upon submission of the consent form, participants were directed to complete the Overall Depression Severity and Impairment Scale (ODSIS) and the Overall Anxiety Severity and Impairment Scale (OASIS) to confirm symptom eligibility at the time of enrollment and to establish baseline depression and anxiety symptom scores. Eligible participants were then randomized, informed they had been successfully enrolled in the study and directed to complete the remaining baseline measures (see **Table 5**, below). Randomization was conducted with a 1:1 allocation of participants using a randomization table with groups determined by random number

generator software. Participants were not directly informed of their condition, although it was not feasible to maintain blinding once support started.

2.8 Follow-Up Evaluation

At 4 weeks post-intervention, participants were emailed a link to complete follow-up survey measures of their depression and anxiety symptoms.

2.9 Retention

We aimed to maximize study retention by reimbursing participants for completion of the intervention, study evaluations and follow-up measures. Compensation was based on how much of the intervention participants completed, and maximum compensation was $100 per participant. Participants who completed the baseline measures and at least 2 weeks of the intervention received $60. Those who completed baseline measures and 11-12 total weeks of the intervention received an additional $20, for a total of $80. Participants who completed both the baseline and 1-month follow-up survey measures were compensated an additional $20, regardless of how many weeks of the intervention they completed. All compensation was provided via electronic gift cards delivered by email. Participants were compensated after the deadline to complete one-month follow-up measures.

2.10 Data Collection

Study evaluation occurred at baseline, weekly throughout the intervention, and at 1-month post-intervention. All data were collected via the study website and completed by participants independently on their own devices. Participants were prompted to complete survey measures by email.

2.11 Outcome Measures

We collected data on experiential engagement, technological engagement, behavioral engagement and depression and anxiety symptoms. We used multiple measures of each facet of engagement, including self-report, behavioral and objective data.[34] Additionally, we used a combination of one-time and repeated measures in order to more comprehensively investigate differences in engagement over time and provide guidance on how researchers might best operationalize engagement in future studies. See **Table 5** for full assessment schedule.

2.11.1 Experiential Engagement

System Usability Scale (SUS). The System Usability Score is a 10-item, validated self-report user experience survey for digital systems answered on a five-point Likert scale.[52] SUS scores range from 0 to 100, with higher scores reflecting more favorable attitudes. The SUS is one of the most commonly used measures of experiential engagement.[15] SUS scores range from 0 to 100, with higher scores representing more positive attitudes. Research indicates that scores above 68 represent above-average attitudes, while scores below 68 represent below-average attitudes. Participants completed the SUS at baseline (to assess anticipatory experiential engagement, before beginning the intervention) and immediately after the final week of the intervention.

Willingness to Refer. Participants were asked at baseline and immediately after every week of the intervention if they would recommend the intervention to others, with yes/no response options.

Referrals. Participants who answered "yes" when asked about their willingness to refer others to the intervention were given the opportunity to provide names and

contact information of any specific individuals they would like to refer to the intervention. We computed the total number of referrals made over the course of the intervention for each participant as a more behavioral measure of experiential engagement.

2.11.2 Technological Engagement

Weekly Completion. The intervention website automatically tracked participants' progress through each week's content. A week was considered started if the participant clicked the emailed link to open the weekly lesson. We calculated weekly completion as the number of responses to weekly prompts (all in-session practice exercises and the writing exercise after the lesson) for which the participant entered any data. Weekly completion did not include any survey measures or homework completion (which we conceptualized as behavioral engagement, see below). From this data, we computed the average weekly completion rate by dividing the number of responses a participant gave that week by the total number of responses possible, resulting in a value between 0 and 100% for each week. This data captures the intensity aspect of technological engagement.[53]

Days Used. The intervention website automatically tracked the date and time when any data was submitted. From this, we computed the number of days each week (0 to 7) when participants submitted any data. We included data related to symptom surveys, weekly lessons, and homework. We did not include other survey measures (which were exclusively for data collection) in our calculations of days used per week. This data, which captures the frequency aspect of technological engagement, distinguishes participants who complete large portions of the intervention content at once, more akin to a traditional therapy session, from participants who complete small portions of the intervention over a longer period of time.[53]

Non-Use Attrition. We operationalized non-use attrition as the last point in the intervention period when any data was submitted. Each participant was assigned a value ranging from 0 to 12, indicating the last week for which any data was submitted. This data captures the time/duration aspect of technological engagement.[53]

2.11.3 Behavioral Engagement

Homework Completion. We conceptualized homework completion as a measure of behavioral engagement, rather than technological engagement, because it entails practicing more adaptive behavioral responses in "real life," outside of the immediate context of the intervention. All written homework assignments were completed on the intervention website. As with weekly completion, we calculated the number of responses to homework prompts divided by the total number of possible responses, resulting in a value between 0 and 100% for each week with homework.

Cognitive Flexibility Inventory. Increased cognitive flexibility is one of the primary mechanisms of change in the Unified Protocol. Change in participants' cognitive flexibility was measured using the Cognitive Flexibility Inventory (CFI). The CFI is an empirically supported, validated 20-item measure of cognitive flexibility designed to be used for repeated measures over the course of an intervention.[54] All items are answered on a 7-point Likert scale with answers ranging from "strongly disagree" to "strongly agree." Scores for each item are summed (some items are reverse scored), with higher scores indicating greater cognitive flexibility. Participants were administered the CFI at baseline and 12 weeks; participants who did not complete Week 12 of the intervention were encouraged to complete this measure when they were contacted for one-month follow-up.

The CFI can be interpreted in three ways: by the total score, by the Alternatives subscale (which reflects the ability to make multiple interpretations of a situation and generate multiple possible solutions) and by the Control subscale (which reflects an internal locus of control related to difficult situations). Total CFI scores can range from 20 to 140; scores on the Alternative subscale range from 13 to 91; and scores on the Control subscale range from 7-49. Higher scores on all three scales reflect greater cognitive flexibility and more adaptive responses to problems.

Cognitive and Affective Mindfulness Scale-Revised. Increased mindfulness is one of the primary mechanisms of change in the Unified Protocol. Change in participants' mindfulness was measured using the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). The CAMS-R is a 10-item questionnaire that assesses mindfulness of thoughts and emotions.[55] All items are answered on a 4-point Likert scale with answers ranging from "rarely/not at all" to "almost always." Scores for each item are summed (one item is reverse-scored). Sum scores range from 12 to 48, with higher scores reflecting greater mindfulness abilities. Participants were administered the CAMS-R at baseline and 12 weeks; participants who did not complete Week 12 of the intervention were encouraged to complete this measure when they were contacted for one-month follow-up.

COPE Inventory - Avoidance Factor. Decreased avoidance is one of the primary mechanisms of change in the Unified Protocol. Change in participants' avoidance behaviors was measured using a 16-item subsection of the COPE Inventory. The COPE Inventory is a survey of coping strategies with 60-items split across 15 4-item scales; there is no summary score. Factor analysis has identified a 16-item avoidance factor,

comprised of four 4-item subscales (mental disengagement, denial, behavioral disengagement, and substance use).[56] To limit response burden and because the remaining 44 items of the COPE Inventory are not related to primary mechanisms of change in the Unified Protocol, we administered only the 16 items of the avoidance factor. Scores on the avoidance factor of the COPE Inventory range from 16-64, with higher scores reflecting greater avoidance. Participants were administered these items at baseline and 12 weeks; participants who did not complete Week 12 of the intervention were encouraged to complete this measure when they were contacted for one-month follow-up.

2.11.4 Depression

Overall Depression Severity and Impairment Scale (ODSIS). Depression symptoms were measured using the Overall Depression Severity and Impairment Scale.[57] The ODSIS was administered at baseline, weekly during the intervention, and at one-month post-intervention.

2.11.5 Anxiety

Overall Anxiety Severity and Impairment Scale (OASIS). Anxiety symptoms were measured using the Overall Anxiety Severity and Impairment Scale.[58] The OASIS was administered at baseline, weekly during the intervention, and at one-month post-intervention.

Table 5. Assessment schedule. Note: Darkened rows indicate outcomes that are automatically collected or derived from the intervention website, rather than measures administered to participants to complete.

		Week													
Assessment	BL	1	2	3	4	5	6	7	8	9	10	11	12	FU

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Demographics	X													
Experiential Engagement														
SUS	X												X	
Willing to Refer		X	X	X	X	X	X	X	X	X	X	X	X	
Referrals		X	X	X	X	X	X	X	X	X	X	X	X	
Technological Engagement														
Completion		X	X	X	X	X	X	X	X	X	X	X	X	
Attrition		X	X	X	X	X	X	X	X	X	X	X	X	
Days Used		X	X	X	X	X	X	X	X	X	X	X	X	
Behavioral Engagement														
HW Completion		X	X	X	X	X	X	X	X	X	X	X	X	
CFI	X												X	
CAMS-R	X												X	
COPE	X												X	
Symptoms														
ODSIS	X	X	X	X	X	X	X	X	X	X	X	X	X	X
OASIS	X	X	X	X	X	X	X	X	X	X	X	X	X	X

2.12 Study Aims

This study aimed to gather more information about the patterns of engagement in this small, exploratory study of the Hazel intervention. We also aimed to assess the feasibility of this intervention and trial design for fully powered RCTs. Specifically, we aimed to: 1) determine if our two arms had sufficient retention and data completion to

allow us to test our hypotheses about differences in engagement between arms; 2) determine if there were differences in attrition based on arm that might require changes to study recruitment or randomization procedures; and 3) assess if support in weeks 8-11 was feasible to examine impacts of this manipulation. We planned to use descriptive statistics, qualitative analyses, and data visualizations to evaluate these questions.[59–64]

2.13 Hypotheses

We hypothesized that there would be no differences between participants on any engagement or health outcomes at baseline due to random allocation. We hypothesized that at post-intervention, participants in the supported arm would show greater experiential engagement than participants in the unsupported arm as a result of the presence of human support. Similarly, we anticipated that there would be few differences in technological engagement between arms for the first seven weeks of the intervention, while all participants were receiving the same intervention and were still blinded to their condition. We hypothesized that starting in Week 8, participants in the supported arm would show greater technological engagement, in line with the model of supportive accountability. As a result of the hypothesized greater experiential and technological engagement in the Supported arm and in line with our conceptual model, we hypothesized that these participants would also show greater behavioral engagement than participants in the unsupported arm. Similarly, we hypothesized that participants in the supported arm would show greater reductions in depression and anxiety symptoms over the course of the intervention, including at post-intervention and 1-month follow-up.

2.14 Statistical Analysis

We planned to use primarily descriptive statistics and data visualizations rather than inferential statistics or p-value significance testing, in line with best practices for pilot and feasibility studies.[59–64]

For all outcomes (demographic characteristics, recruitment results, engagement, and depression/anxiety symptoms), we computed descriptive statistics including mean, standard deviation, range and 95% confidence interval for continuous variables; n and percentage for categorical variables. For variables collected twice over the study period (e.g., baseline and 12 weeks), we present pre/post descriptives and visualizations. For variables collected weekly, we present averages and visualizations of trends over time by both timepoint and by participant.

We decided a priori to present experiential engagement outcomes by arm, as we hypothesized that the presence/absence of support which might lead to meaningful differences in the subjective experience of using the intervention. We similarly decided a priori to present technological engagement outcomes by arm, as our experimental manipulation involved different technology (self-guided study website versus self-guided study website plus video sessions) which could feasibly result in different usage patterns. We did not make an a priori decision about how to present behavioral engagement and health outcomes. Rather, based on our conceptual model (see **Figure 1**), we planned to decide post-hoc how to present these outcomes based on the results of our experiential and technological engagement outcomes. Specifically, if there were potentially meaningful differences between arms on experiential and/or technological engagement outcomes, we planned to present behavioral engagement and health outcomes by arm as well. However, if there were few to no potentially meaningful

differences between arms on these upstream facets of engagement, we planned to present behavioral engagement and health outcomes for the sample as a whole.

All data were analyzed using R 4.0.0 and RStudio 1.2.5042. All data visualizations were created using Microsoft Excel 16.54 and Adobe Illustrator 25.4.5.

Chapter 3. Results

Our study received funding on July 30, 2020. Ethical approval was obtained on March 15, 2021. Recruitment began in April 2021 and concluded in December 2021. Data collection occurred between April 2021 and March 2022.

3.1 Participants

Participants were recruited between April 7, 2021, and December 31, 2021. The online screener was completed 317 times by 275 unique individuals. Of these, 70 were potentially eligible and were contacted for phone screening. 38 were unable to be reached for screening, 8 were ineligible based on phone screening, 0 were eligible but declined to enroll, and 26 individuals were enrolled in the study (13 per am). The most common reasons that respondents were found to be ineligible from the online screening were due to taking psychiatric medication or having insufficient symptoms of depression and anxiety. The most common reason respondents were found ineligible on the phone screening was due to insufficient symptoms at baseline (n=6). 3 enrolled participants withdrew from the study during the intervention (one stated they had not benefited from the study; one stated it was not the right fit; and one did not provide a reason); 23 participants were included in the following analyses. One withdrawn participant was in the self-guided arm and two were in the supported arm; however, all three withdrew before study therapy sessions began and while they were still blinded to condition. See **Figure 3** for CONSORT Diagram.

Notably, four additional participants emailed the PI reporting they were unable to continue in the study (two due to being too busy, two due to deciding to seek in-person therapy during the study window) yet elected to remain enrolled so their available data would be included in analyses. Two of these participants were in the self-

guided arm and two were in the supported arm; all stopped participating before study therapy sessions would have begun.

3.1.1 Effectiveness of Different Recruitment Strategies

Of the 23 participants we enrolled, the majority were recruited from ResearchMatch (20; 86.96%). One participant was recruited from a different professional organization (Duke Clinical Trials Registry; 4.35%), one was recruited through word-of-mouth from another participant (4.35%), and one was recruited from social media (Instagram; 4.35%).
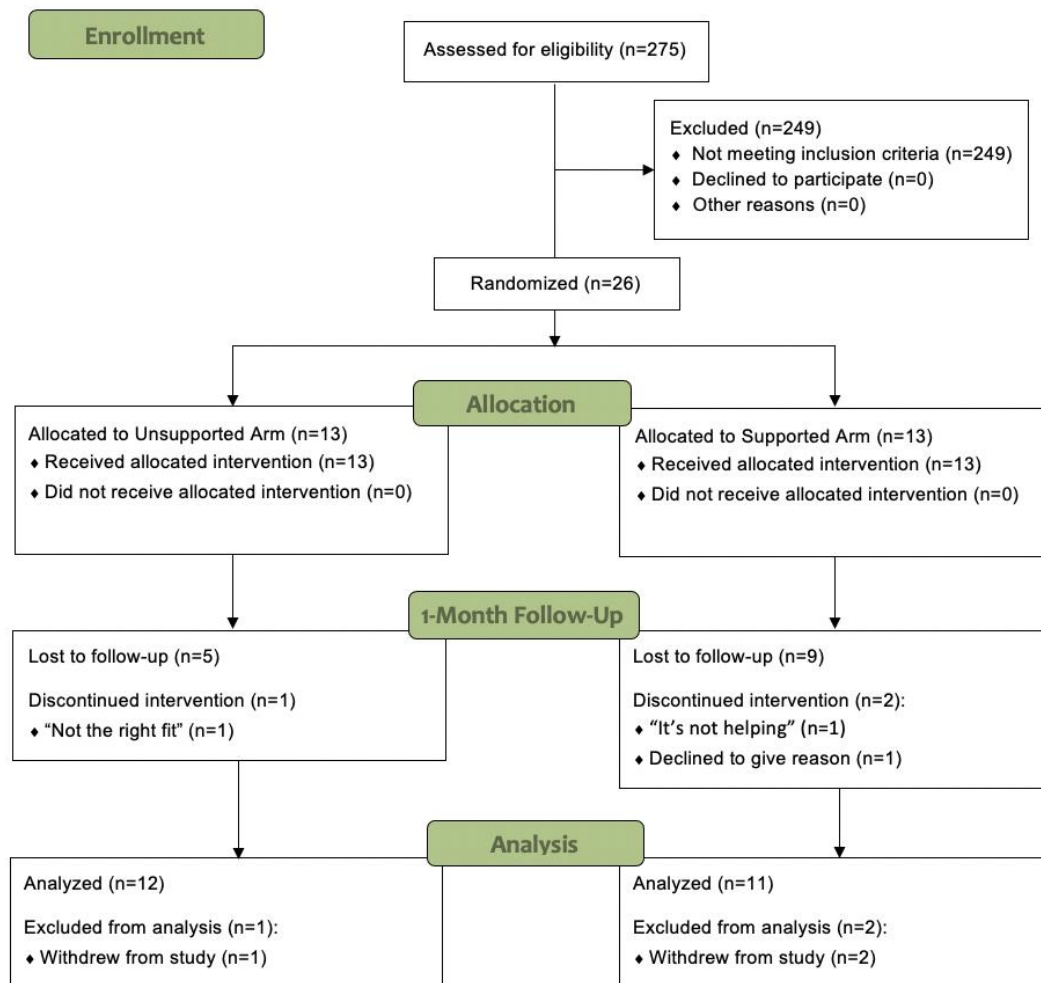
Figure 3: CONSORT Diagram.

3.1.2 Sample Demographic Characteristics

Participants ranged in age from 20 to 66 years. The average age was 38 years. A minority of participants identified as BIPOC (17%) or members of the LGBTQ+ community (21%). Slightly more than a third of the sample (35%) had educational attainment of less than an associate's degree. No participants self-identified as Hispanic or Latinx. No participants self-identified as multiracial.

Table 6. Sample demographic characteristics.

	N (%)
Age ($\bar{x}$, SD)	37.89 (12.07)
Gender Identity	
Cisgender Woman	20 (86.96%)
Cisgender Man	2 (8.70%)
Transgender Woman	1 (4.35%)
Education Level	
Less than HS diploma	1 (4.35%)
HS or equivalent	6 (26.09%)
Some college	1 (4.35%)
2-year degree	1 (4.35%)
4-year degree	4 (17.39%)
Some grad school	2 (8.70%)
Graduate degree	8 (34.78%)
Sexual Orientation	
Heterosexual	19 (82.61%)
Gay/Lesbian	2 (8.70%)
Bisexual	2 (8.70%)

Race	
White	19 (82.61%)
Black/African American	2 (8.70%)
Native Hawaaiian/Pacific Islander	2 (8.70%)

3.2 Experiential Engagement: How did participants feel about the intervention?

For our experiential engagement outcomes, we conducted all analyses by arm to assess for signs of differential attitudes toward the intervention based on the presence/absence of therapist support. We hypothesized that there would be minimal differences between arms at baseline and, for measures of experiential engagement collected weekly, for the first 7 weeks of the intervention while participants remained blinded to condition. Participants in the supported arm became aware of this in Week 8 when their study therapist was introduced, and we anticipated that differences in attitudes toward the intervention might begin at this point and be maintained for the remainder of the intervention. Results for each outcome are presented below.

3.2.1 Intervention Usability

At baseline, participants in the unsupported arm (n=11) had a mean System Usability Scale score of 76.82 (SD=18.51). Participants in the supported arm (n=11) had a mean baseline SUS score of 74.09 (SD=16.14). Scores in the mid-70s represent overall positive attitudes towards an intervention. See **Table 7** for full results. At 12 weeks, participants in the unsupported arm (n=8) had a mean SUS score of 68.44 (SD=16.74), and participants in the supported arm (n=4) had a mean score of 71.88 (SD=29.68). This

represents a 2.2-point decrease in mean SUS score for the supported arm over the course of the intervention, and an 8.4-point decrease in mean SUS score for participants in the unsupported arm.

Table 7. Intervention usability at baseline and post-intervention, by arm.

	Unsupported				Supported			
	X̄ (n)	SD	Range	95% CI	X̄ (n)	SD	Range	95% CI
Baseline SUS	76.81 (11)	18.51	40 - 100	±10.94	74.09 (11)	16.14	52.5-100	±9.54
12-Week SUS	68.44 (8)	16.74	40 - 92.5	±11.60	71.88 (4)	29.68	30-100	±29.08

3.2.2 Willingness to Refer

Overall, participants were more likely to answer "yes" when asked if they would recommend the intervention at all but two time points (see **Figure 4** for more). Slightly over half of all participants (n=12) never changed their answer to this question across all time points for which they submitted data. Of these, 10 participants answered "yes" at every time point for which they submitted data, compared to 2 participants who answered "no" at every time point. The remaining 11 participants gave both answers at different time points. Of those 11, 4 more often said "yes", 6 more often said "no", and one gave both answers an equal number of times. Of the same 11, 4 initially answered "yes" on at least one time point but switched to consistently answering "no" at subsequent time points. None who initially answered "no" changed to consistently answering "yes," and the remaining 6 had multiple changes between "yes" and "no" responses across the intervention window. Participants were most likely to answer "yes" in weeks 1 (78.26%, n=23), 2 (80.00%, n=20), and 8 (80.00%, n=10). Participants were

most likely to answer "no" in weeks 4 (44.44%, n=9) and 5 (45.45%, n=11). Scores generally decreased over time, with boosts in weeks 8 and 12; notably, these were the two weeks with no homework assignment.

Contrary to our expectations, there were some differences in willingness to refer between arms at baseline. Participants in the unsupported arm were more likely to be willing to recommend the intervention (83.33%) than participants in the supported arm (72.72%). Further, 7 of the 10 participants who answered "yes" at all timepoints were in the unsupported arm, as well as one participant who said "yes" at 11 of the 12 timepoints. Only 3 participants in the supported arm said "yes" at all time points. Participants in the supported arm were less willing to recommend the intervention at all time points except for the final week of the intervention, when 75% in the supported arm answered "yes" compared to 63% in the unsupported arm. Notably, no participants in the supported arm answered this item for weeks 9 or 11, and in Week 10 only one participant responded.
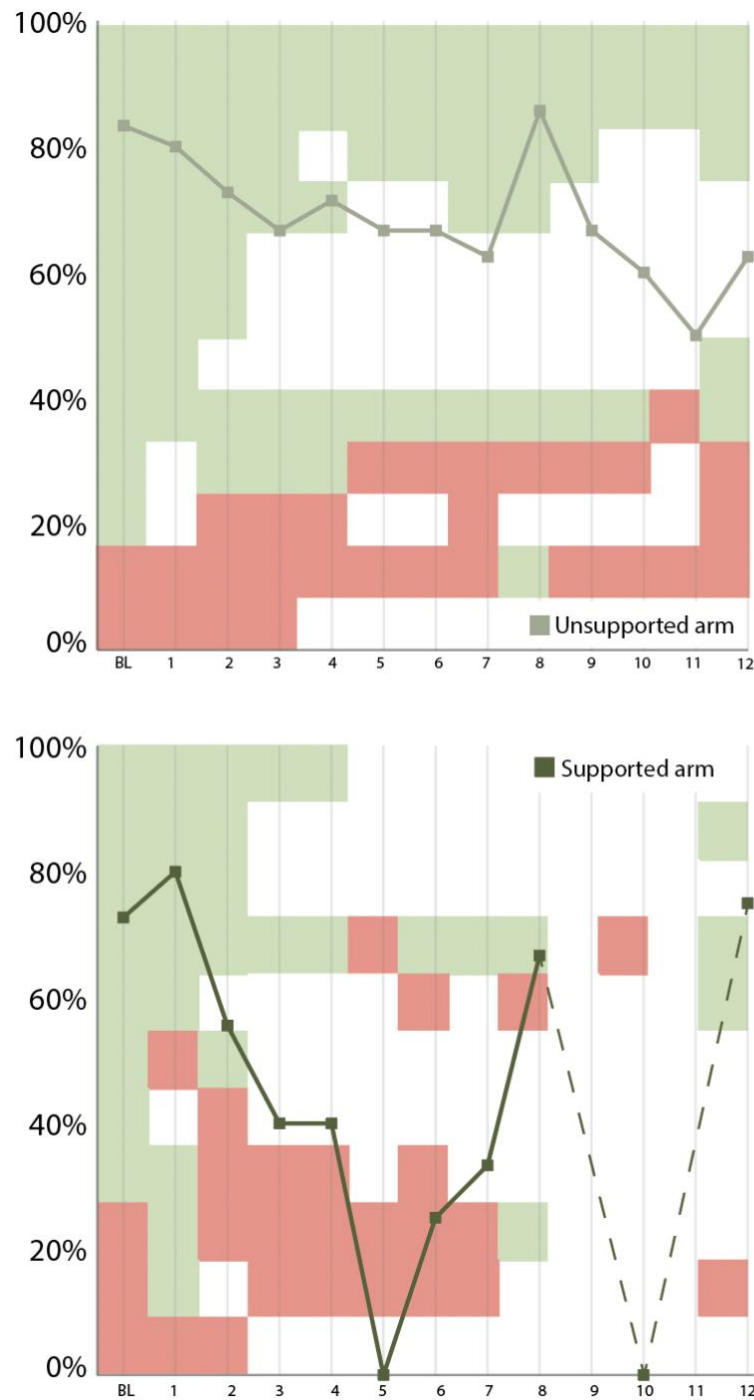
Figure 4. Willingness to recommend the Hazel intervention (%) per week, by arm, plus individual participant response patterns. Line graph represents the percentage of respondents each week who answered "yes" when asked if they would recommend the Hazel intervention to others. Green, white and red background represents individual participant responses each week (read horizontally); green represents answer of "yes"; red represents "no"; white represents no answer.

44

3.2.3 Number of Referrals

Only 3 of the 23 participants made a referral at any point in time. Two participants from the unsupported arm (16.67%) and one participant from the supported arm (9.09%) made referrals. The participants in the unsupported arm made one referral each. The participant in the supported arm made a total of two referrals.

3.3 Technological Engagement: How did participants use the intervention?

For our technological engagement outcomes, we conducted all analyses by arm to assess for signs of differential patterns of use based on the presence/absence of therapist support. We hypothesized that the different delivery modality between arms might lead to differing patterns of use, especially in weeks 8-11 when video therapy sessions occurred. Per the model of supportive accountability, we hypothesized that participants in the supported arm might show a boost in technological engagement during these weeks. Results for each outcome are presented below.

3.3.1 Weekly Lesson Completion

To understand how participants used the intervention, we sought to parse apart those who fully completed the weekly lessons from those who partially completed lessons and those who never started lessons. To do so, we first calculated the percentage of participants who *started* each weekly lesson (see **Figure 5**). At first both arms follow similar patterns, with 100% of participants starting the first two weeks' lessons followed by a decline around weeks 3-5. From Week 4 onward, the unsupported arm levels out for the remainder of the intervention with a relatively consistent number of participants starting each subsequent lesson; there is one notable drop during Week 11 only. For the supported arm, there is also a leveling off after the decline in number of participants starting each lesson during weeks 3-5; however, for this arm there is a second decline in

weeks 9-11. The supported arm also showed a slight increase in Week 12, but even with this increase more participants in the unsupported arm started Week 12.

Notably, differences between arms emerge before Week 8, while participants were still blinded to study condition. There were no differences between arms in the delivered intervention during weeks 3-7 which might account for lower lesson initiation in the supported arm. The difference between arms becomes most extreme during weeks 9-11, overlapping with the provision of video therapy sessions for participants in the supported arm. **Figure 6** reflects the number of video therapy sessions completed by participants in this arm. Most (n=7) did not complete any video therapy sessions, and only 1 of the 11 participants in this arm completed all 4 video sessions.
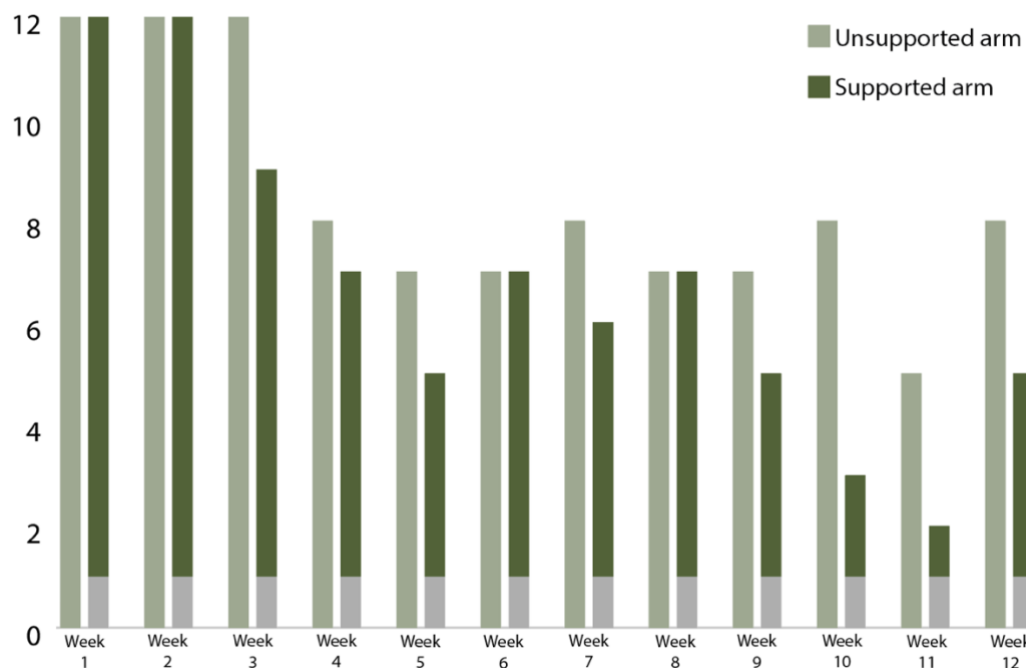

Figure 5: Participants starting each weekly lesson by arm. Note that the supported arm is boosted by one participant (gray bars) every week to allow for easier comparison between arms given uneven allocation (n=12 unsupported versus n=11 supported).
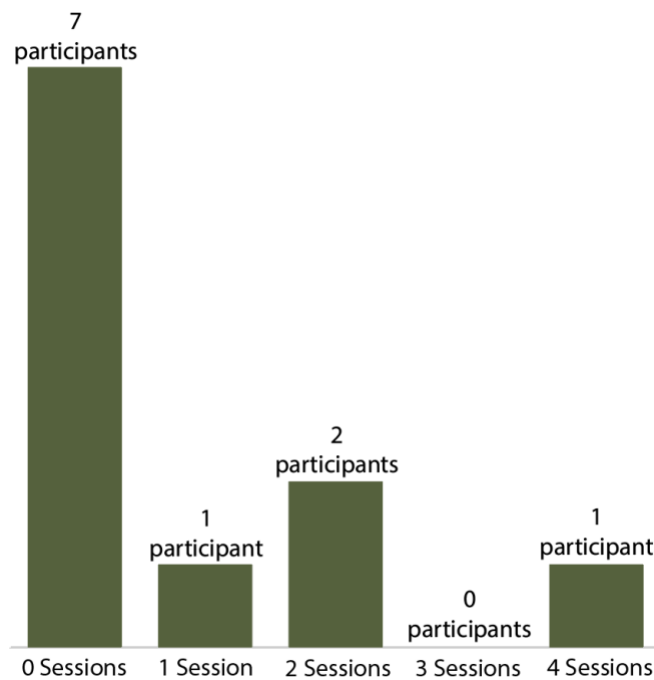
Figure 6. Total number of video therapy sessions completed by participants in the supported arm.

Next, we calculated participants' percent completeness of each weekly lesson. We did so by dividing the number of responses that a participant gave in each lesson by the total number of responses possible for that lesson. Participants who never started the intervention were excluded from our subsequent analyses of weekly completion. We excluded those participants as their noncompletion was already captured by our previous analyses and to prevent meaningful information about participants' behavior after starting each weekly lesson from being washed out by the high numbers of non-starters.

Among participants who at least started each weekly lesson, there was relatively high completion for much of the intervention, with mean completion of above 70% in both arms for the first 6 weeks of the intervention. There were no major differences between arms until weeks 10 and 11, at which point participants in the supported arm had much higher mean completion scores than participants in the unsupported arm — with a

47

difference of 29.5% between arms in Week 10 and 66.7% in Week 11. See **Figure 7** for more.
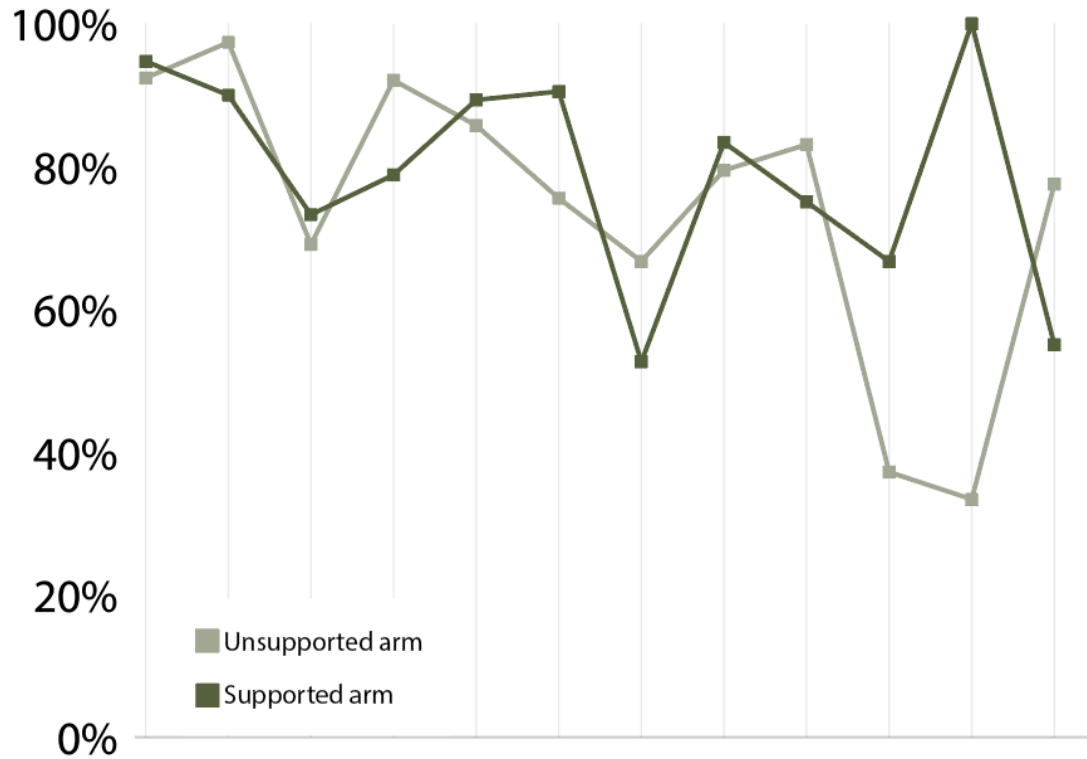
Figure 7. Average lesson completeness (%) per week by arm.

3.3.2 Non-Use Attrition

We calculated the point of non-use attrition as the last week for which participants submitted any data. Given the low sample size, we have considered attrition of 2 or more participants per arm per week as potentially meaningful spikes. See **Figure 8** for new and total attrition each week.

There were minimal differences in attrition by arm for the first 7 weeks of the study; this was expected given minimal differences between arms for this time period. For both arms, there was a spike in attrition in Week 3, with 5 participants across both arms dropping out between weeks 2 and 3. A second spike in attrition occurred in weeks

48

8 and 9 of the intervention, for the supported arm only. Across these two weeks, 4 participants in the supported arm were lost to follow-up, compared to one participant in the unsupported arm. Notably, this spike corresponds with the onset of video sessions for the supported arm.
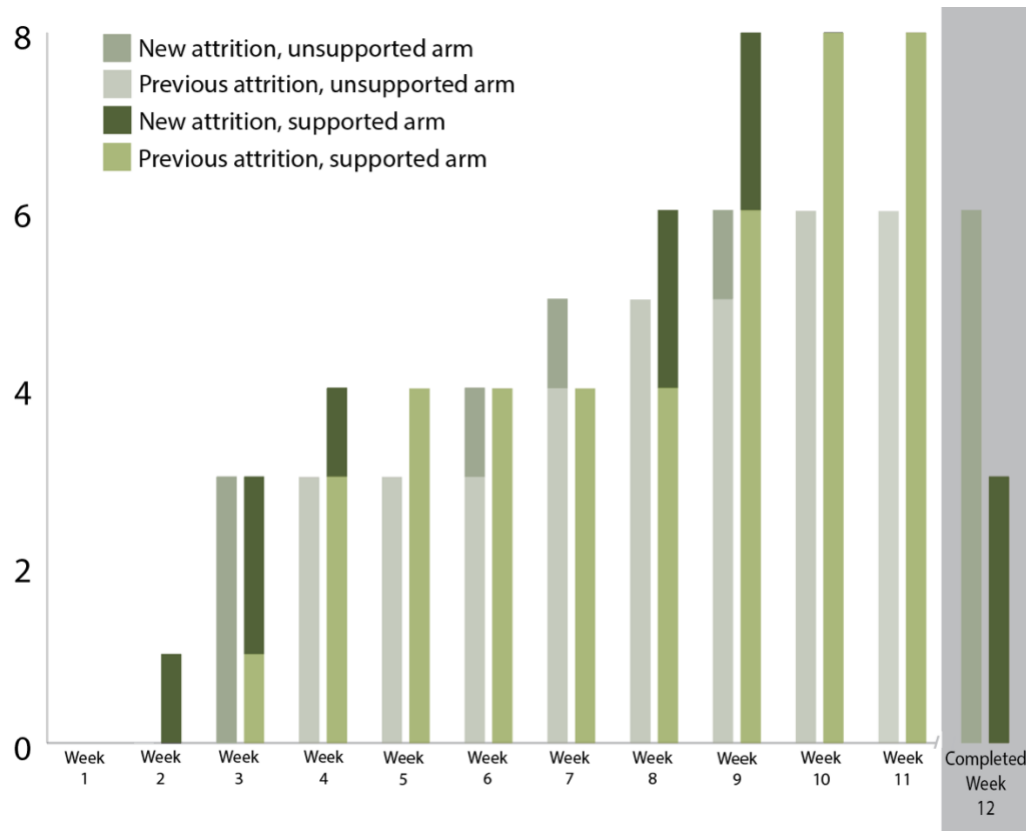
Figure 8. New and total attrition per week by arm.

3.3.3 Days Used Per Week

Overall, there was minimal variability in how many days per week participants used the intervention, with no noticeable differences between arms. Participants in the unsupported arm used the intervention for an average of 1.309 days per week (SD=.667), compared to an average of 1.227 days per week for participants in the supported arm (SD=.549). Further, of our 23 participants, just over half (n=12) exclusively used the intervention for one day per week; these 12 were evenly divided between the two arms.

49

The maximum number of days the intervention was used in a given week was 4. The

participant with the highest dose of exposure to the intervention used it for an average of

2.00 days per week across all intervention weeks; this participant was in the unsupported

arm. Week 2 involved the lowest dose of use for all weeks; all participants who

completed Week 2 used it for a single day only. Week 5 was used the most days, with an

average of 1.636 days per week.

3.4 Behavioral Engagement: How did participants' behavior change over the course of the intervention?

For our behavioral engagement outcomes, we conducted all analyses for the

sample as a whole rather than by arm given the low number of video therapy sessions

completed by participants in the supported arm. The low number of video sessions means

there were minimal differences in the delivered intervention between arms, so we

anticipated minimal differences in these outcomes between arms. Results for each

outcome are presented below.

3.4.1 Weekly Homework Completion

As with weekly lesson completion (see above), we calculated the percent

completeness of each week's homework (note that weeks 8 and 12 had no homework

assigned) for each participant that at least *started* that week's lesson. Again, we excluded

participants who did not start the weekly lesson from these calculations to avoid

artificially deflating our averages and losing potentially meaningful information about

how participants who were exposed to each new weekly skill incorporated that skill into

their day-to-day lives.

Among participants who at least started each weekly lesson, there was high mean homework completion for Week 1 (83.23%). Homework completion was low for the remainder of the intervention, with Week 6 as the only other week where mean homework completion was above 50% (58.135%). Week 3 had the third-highest mean homework completion (45.71%); for all other weeks, participants who started the lesson completed less than 30% of the homework on average (see **Figure 9** for more).
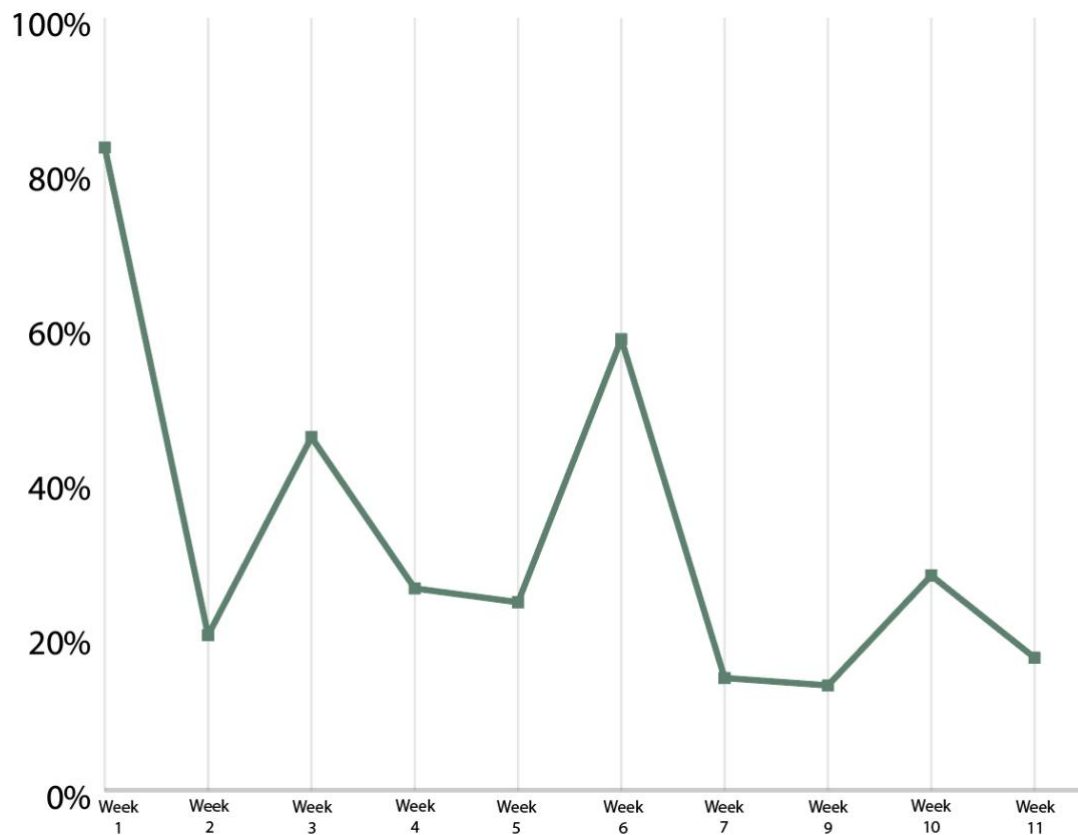

Figure 9. Average homework completeness (%) per week.

3.4.2 Cognitive Flexibility

At baseline, participants (n=23) had an average total CFI of 100.826 (SD=14.908), indicative of relatively high baseline cognitive flexibility. Participants had a mean baseline score of 72.087 (SD=9.429) on the alternatives subscale and a mean

51

baseline score of 28.739 (SD=8.058) on the control subscale, see **Table 8** for more. Relative to the maximum score for each subscale, participants scored higher on the alternatives subscale (80.00%) compared to the control subscale (58.65%).

After completing the intervention, participants (n=12) had a mean total CFI of 95.917 (SD=12.926), a mean alternatives subscale score of 67.000 (SD=7.966) and a mean control subscale score of 28.917 (SD=8.382). Notably, there was an approximately 5-point decrease in the mean total CFI score over time. From comparing pre/post scores on the two subscales, we see that this decrease appears to be driven entirely by decreases in the alternatives subscale, as mean control subscale score shows a very minor increase (0.178).

Table 8. Behavioral engagement outcomes at baseline and post-intervention.

	X̄(n)	SD	Range	95% CI
Baseline CFI	100.826 (23)	14.908	75 - 136	± 6.093
Alternatives Subscale	72.087	9.429	55 - 91	± 3.853
Control Subscale	28.739	8.058	18 - 45	± 3.293
12-Week CFI	95.917 (12)	12.930	74 - 114	± 7.316
Alternatives Subscale	67.000	7.966	53 - 77	± 4.507
Control Subscale	28.917	8.382	20 - 42	± 4.742
Baseline CAMS-R	23.870 (23)	2.437	20 - 28	± 0.996
12-Week CAMS-R	24.667 (12)	3.798	20 - 31	± 2.149
Baseline COPE-Avoidance	30.652 (23)	8.205	18 - 48	± 3.353

12-Week COPE-Avoidance	29.417 (12)	5.501	21 - 39	± 3.112

3.4.3 Mindfulness

Participants (n=23) had a mean baseline CAMS-R score of 23.870 (SD=2.437), compared to a mean post-intervention (n=12) score of 24.667 (SD=3.798), see **Table 8** for more.

3.4.4 Avoidance

At baseline, participants (n=23) had a mean COPE-Avoidance score of 30.652 (SD=8.205). At 12 weeks, participants (n=12) had a mean score of 29.417 (SD=5.501), see **Table 8** for more.

3.5 Symptom Change: How did participants' symptoms change over the course of the intervention?

We originally planned to calculate best-fit lines participants' symptom scores over time to reflect the general trend in symptoms. However, after finishing data collection we observed highly variable trajectories of change which contraindicated the use of a singular best-fit line (see **Figure 10** and **Figure 11** for spaghetti plots of depression and anxiety symptoms over time). We still present the average symptoms over time, but have decided post-hoc to additionally divide the sample into categories based on similar trends in symptom change over time to better understand the variability in symptom change. Given the minimal differences between arms on engagement outcomes and the low number of video sessions completed by participants in the supported arm, we have looked at symptom outcomes for the entire sample rather than by arm.
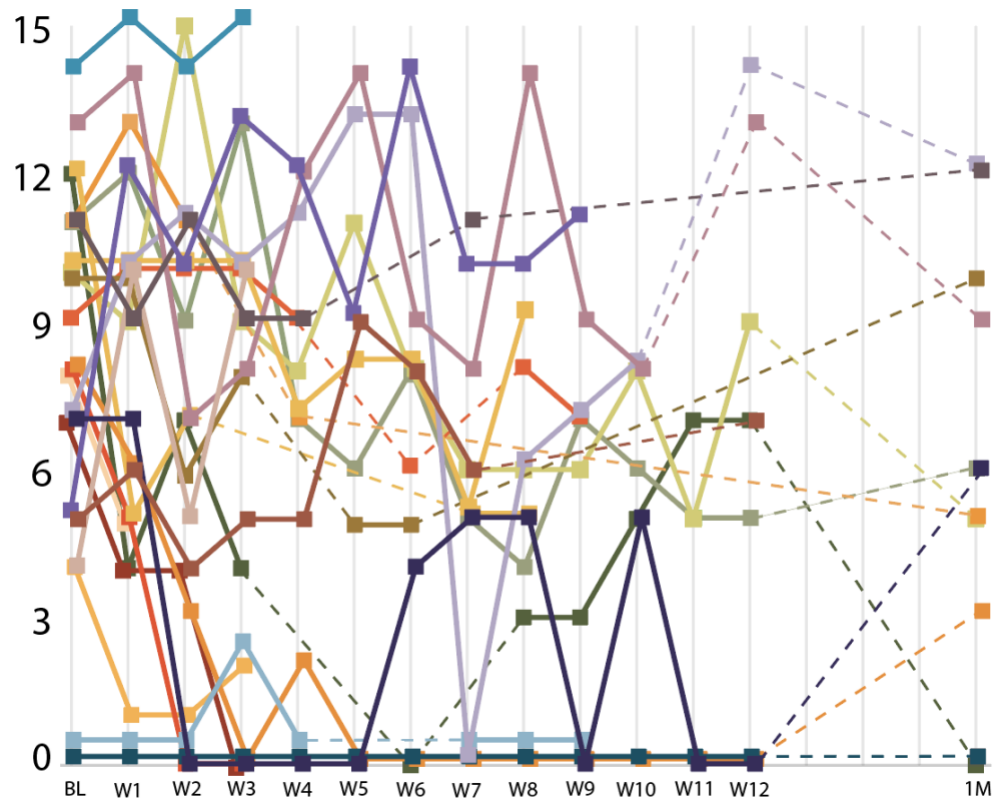
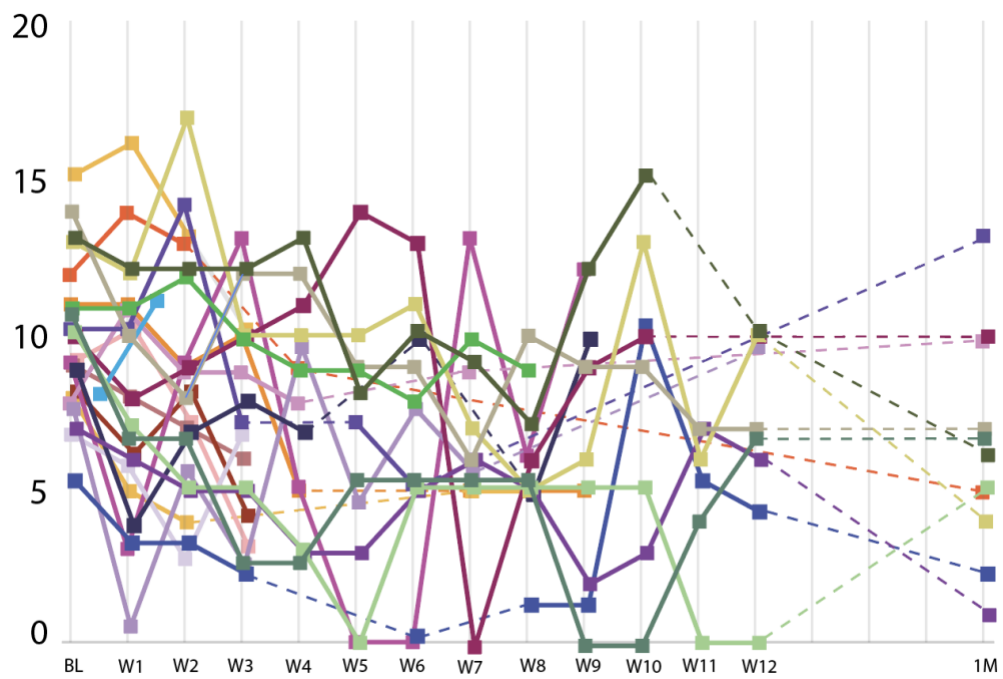**Figure 10. Spaghetti plot of ODSIS scores over time.**

Figure 11. Spaghetti plot of OASIS scores over time.

3.5.1 Depression

Across all participants, symptoms had a general downward trajectory in depression symptoms, as measured by the ODSIS score over the course of the intervention until week 12, where there was a roughly 4-point increase; this increase was roughly sustained at one-month follow-up (see **Figure 12** below). However, as can be seen in the spaghetti plot of ODSIS scores (see **Figure 10** above), there were several participants with minimal depression symptoms (n=2) at all timepoints, which is artificially deflating these averages. There are also several additional participants (n=7) who had subthreshold depression scores at baseline (ODSIS below our cut point of 8). The results in **Figure 12** may be biased by the inclusion of numerous participants without interfering depression symptoms, so we therefore also calculated the average ODSIS score over time for the 14 participants with significantly interfering/impairing depression symptoms at baseline, see **Figure 13** below. This subset of our sample showed a similar trend in ODSIS score over time, with a decrease through Week 11 and then a moderate increase in Week 12 that was sustained at 1-month follow-up. However, the decrease from baseline to Week 11 was larger and more consistent for the depressed-only group, and the increase at Week 12 was smaller (roughly 2 points versus 4). Further, for the depressed-only group, there was a roughly one-point decrease from Week 12 to one-month follow-up, compared to a slight increase for the total sample.
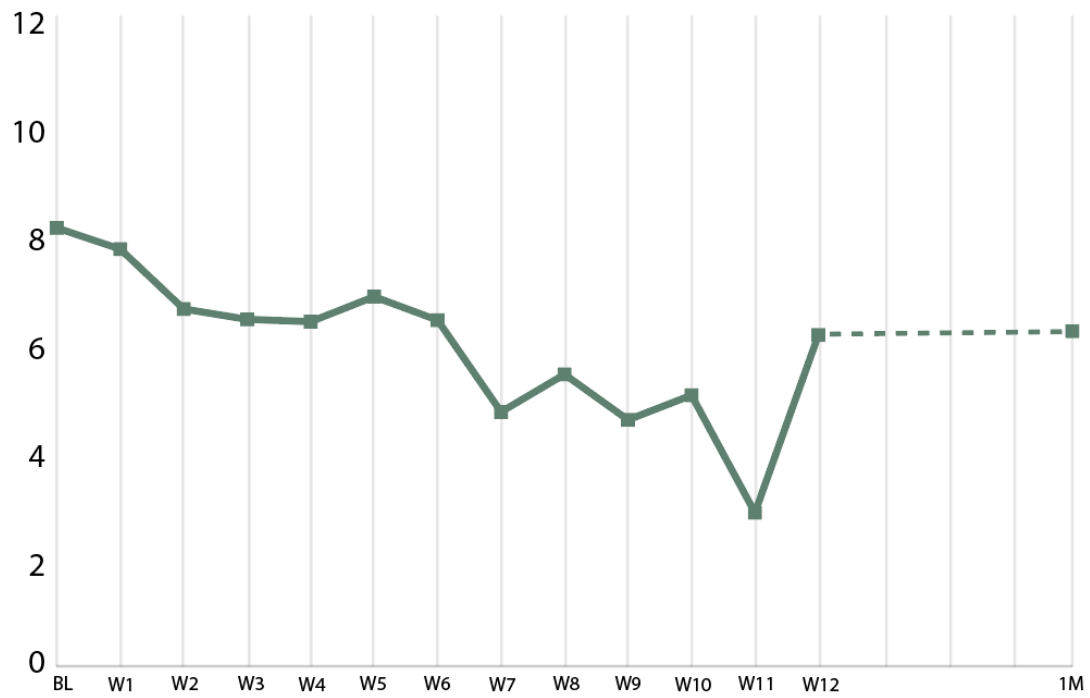
Figure 12. Average ODSIS score over time for all participants (n=23).
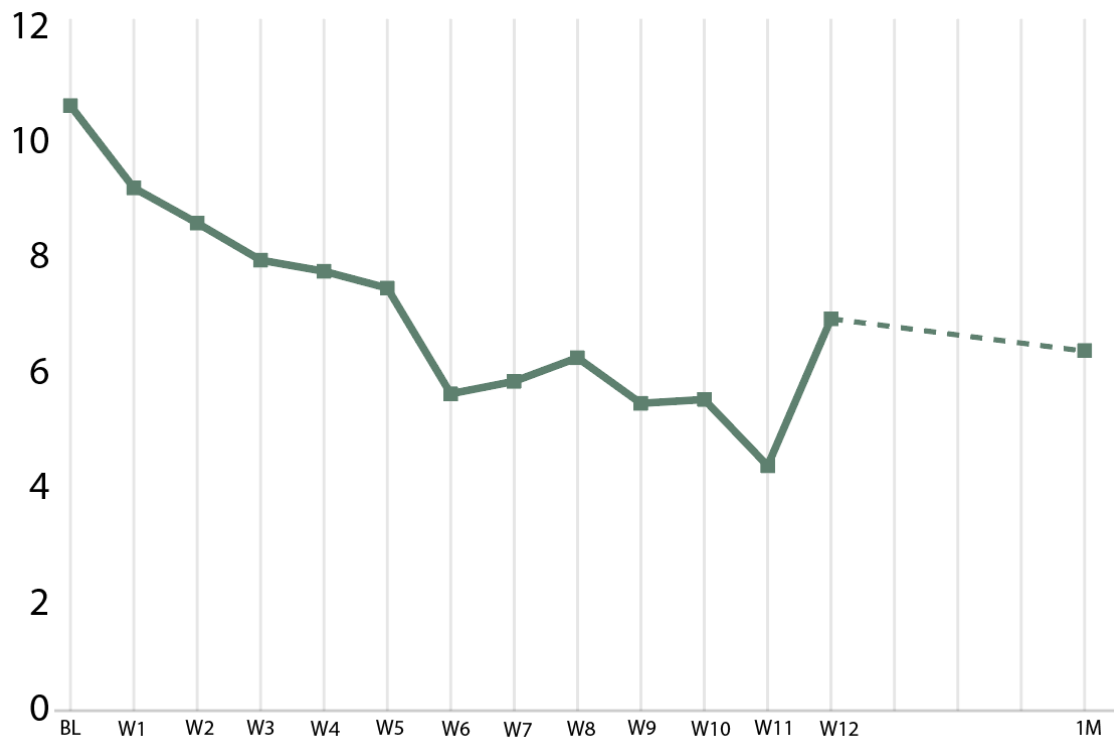

Figure 13. Average ODSIS score over time for participants with above-threshold depression symptoms at baseline (n=14).

We then separated out individual trajectories from the spaghetti plot of ODSIS scores and grouped individuals by the pattern over time. We observed five unique patterns of depression symptom change over the course of the intervention; see **Figure 14** for individual trends. Most participants (n=13) showed some improvement in their symptoms over the course of the intervention, and most (n=8) who showed improvement had relatively consistent downward trajectories over time (see **Figure 14.a**). However, three of the 8 in this category tracked their symptoms for only 3 weeks of the intervention, and one of the 8 tracked for only one week of the intervention.

An additional three participants showed a general reduction in symptoms over time, with multiple spikes and drops at different timepoints (see **Figure 14.b**). This is in contrast to the 8 participants with improved symptoms and no spikes. We conceptualized a symptom spike as an increase of at least 2 points in a week followed by a drop of at least 2 points. For these three participants, symptom increases were consistently less extreme than the subsequent decreases, leading to an overall downward trend in symptoms despite these spikes. Two of the three participants displayed a pattern of frequent spikes often lasting 2-3 weeks each, whereas one showed a longer cycle lasting closer to 5-6 weeks per spike.

An additional two participants showed early improvement in symptoms, followed by an uptick in symptoms at the last time point where data was available (see **Figure 14.c**).

Seven participants showed a cyclical pattern in symptoms with no overall reduction or increase in symptoms from baseline to post-intervention (see **Figure 14.d**). This is in contrast to the 3 participants mentioned above with cyclical, but overall

improving, symptoms. As with those participants, these 7 participants varied in the

frequency of their symptom cycles. Three of these 7 participants had cycles that typically

lasted around 2-3 weeks, while four of the 7 had symptom cycles lasting around 4-6

weeks per cycle.

Finally, three participants showed little change in their symptoms over time (see

Figure 14.e). Notably, two of these three had minimal depression symptoms at all time

points. The third had elevated symptoms (ranging from an ODSIS score of 14 to 15) but

completed only 3 weeks of the intervention.
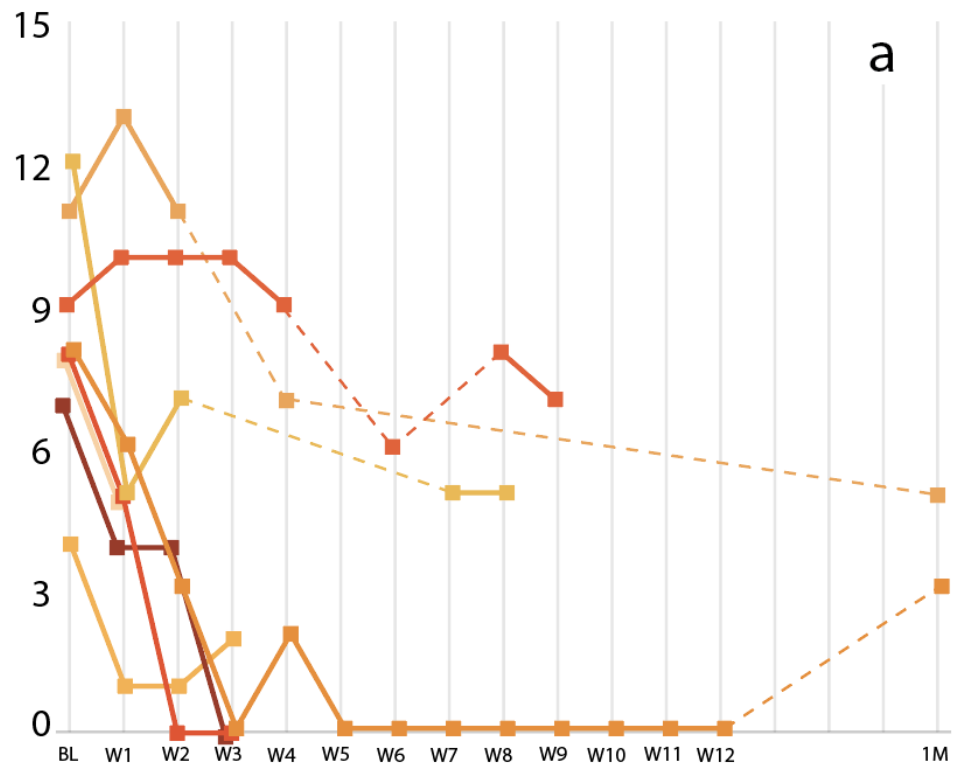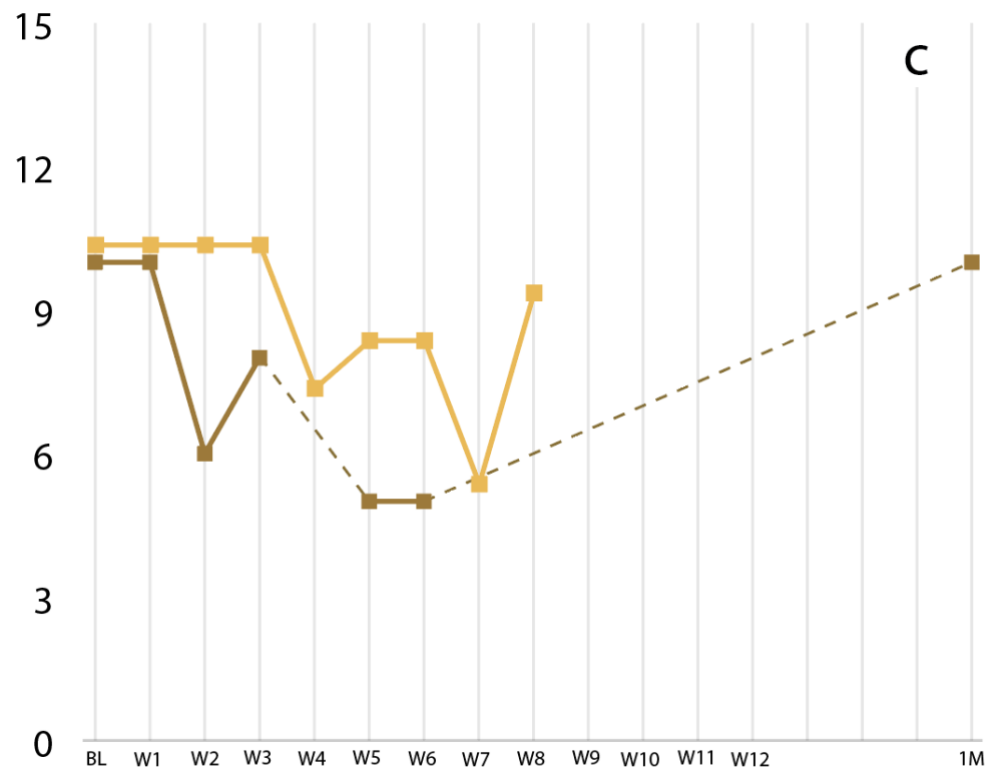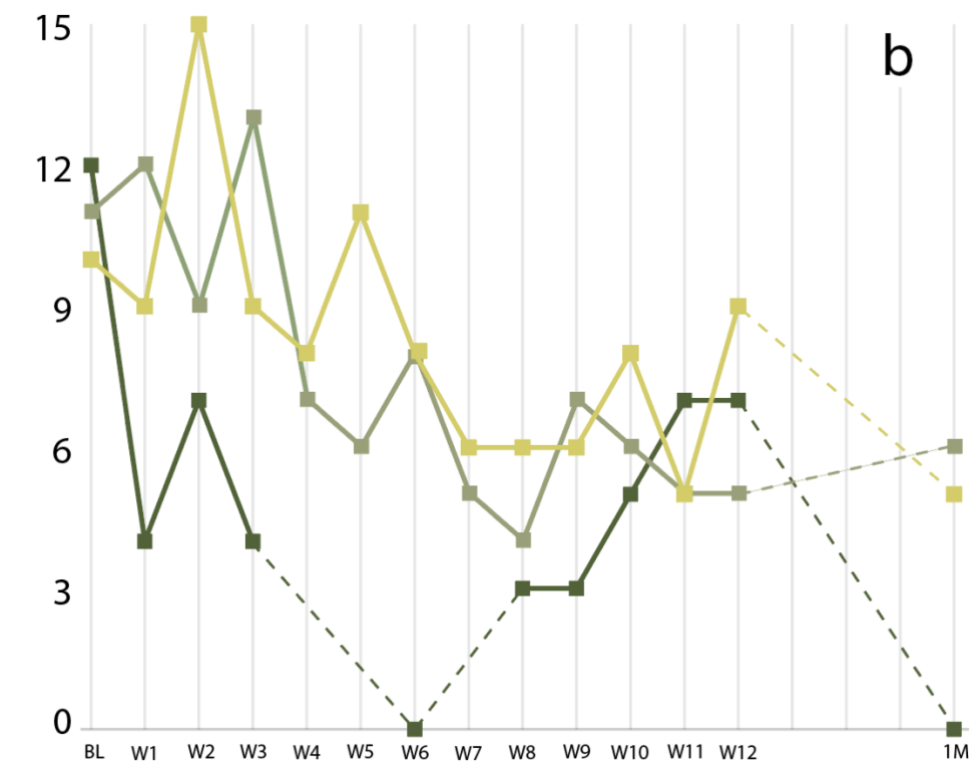
Figure 14. Individual ODSIS trajectories, grouped by pattern. A. Consistent improvement, n=8. B. Improvement with cyclical pattern, n=3. C. Temporary improvement followed by increased symptoms, n=2. D. Cyclical pattern with no improvement, n=7. E. Minimal change, n=2.

3.5.2 Anxiety

As with depression symptoms, there was a general downward trend in average OASIS scores over time, with a notable spike toward the end of the intervention (see **Figure 15** below). Symptoms generally decreased from baseline through Week 8, rose in weeks 9 and 10, dropped in Week 11 and rose again in Week 12. There was a slight decrease in OASIS score from Week 12 to one-month follow-up.

As with depression symptoms discussed above, there were several participants with subthreshold anxiety symptoms at baseline (n=3). As above, we plotted the average OASIS score over time with these participants removed to better understand how the intervention impacted anxiety symptoms for those with significantly impairing/interfering anxiety (see **Figure 16** below). The patterns in OASIS scores after removing these participants was generally similar to that observed in the full sample. Notably, the spike in weeks 9 and 10 was slightly more pronounced in the anxious sample, and the decrease in Week 11 was more extreme as well. Additionally, there was a less-pronounced decrease from Week 12 to one-month follow-up for this subset of participants.
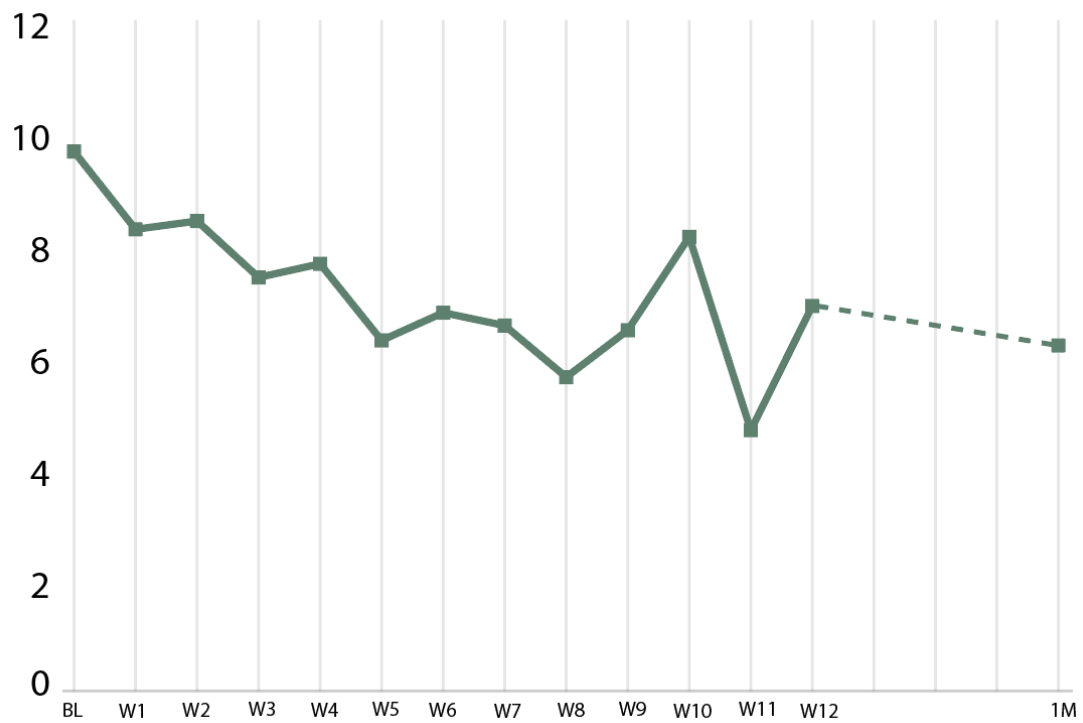
Figure 15. Average OASIS score over time for all participants (n=23).
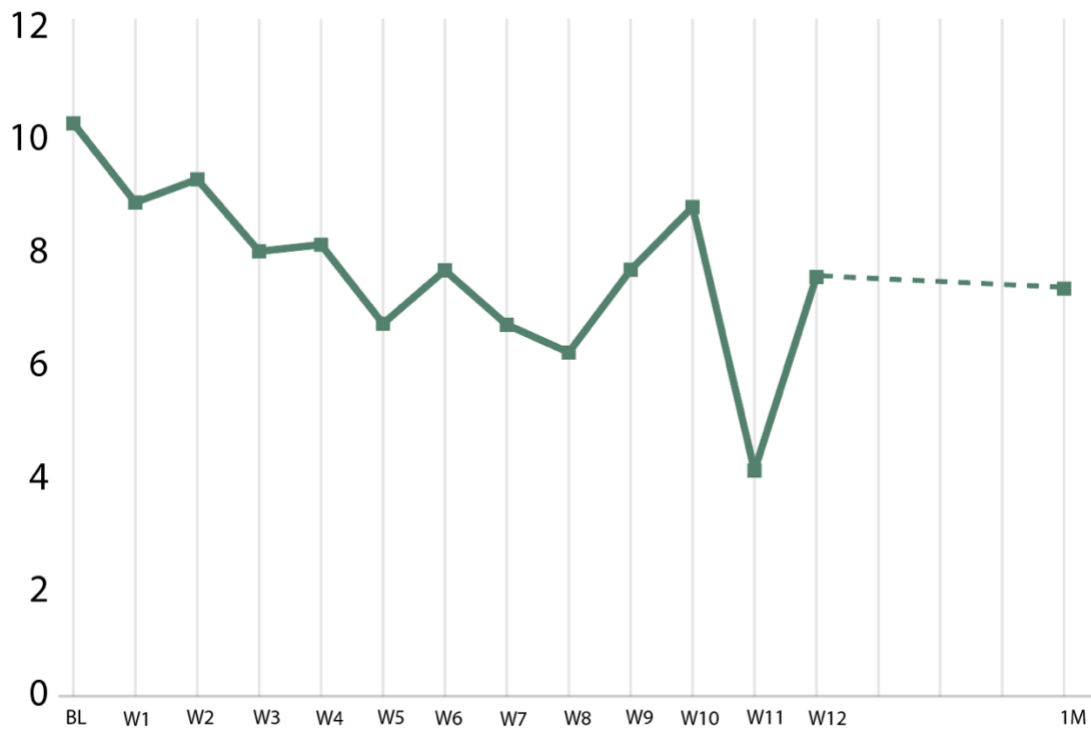
Figure 16. Average OASIS score over time for participants with above-threshold anxiety symptoms at baseline (n=20).

As with depression symptoms, we then separated out individual trajectories from the spaghetti plot of OASIS scores and grouped individuals by the pattern over time. We observed four unique trends in anxiety symptoms over the course of the intervention, see **Figure 17** for individual trends. Most participants (n=13) showed an overall reduction in anxiety symptoms over time. Again, these participants could be split into those who showed generally consistent improvements in symptoms over time (n=7, see **Figure 17.a**) and those who showed a pattern of spikes and drops with an overall downward trajectory (n=6, see **Figure 17.b**). However, four of the 7 who showed consistent improvement completed anxiety symptom monitoring for only 3 weeks, so our ability to draw conclusions is limited by missing data.

An additional 9 participants showed cyclical symptoms with no overall increase or reduction in symptoms from baseline to post-intervention (see **Figure 17.c**). Compared to the cyclical depression symptoms discussed above, these cyclical anxiety symptoms tended to be more unpredictable in terms of their frequency and the extremity of each spike/drop.

Lastly, one participant showed worsening anxiety symptoms over the course of the intervention (see **Figure 17.d**). However, this participant completed symptom tracking for only two timepoints (baseline and Week 1), so we cannot draw any conclusions about how this participant's symptoms truly changed over the intervention period.
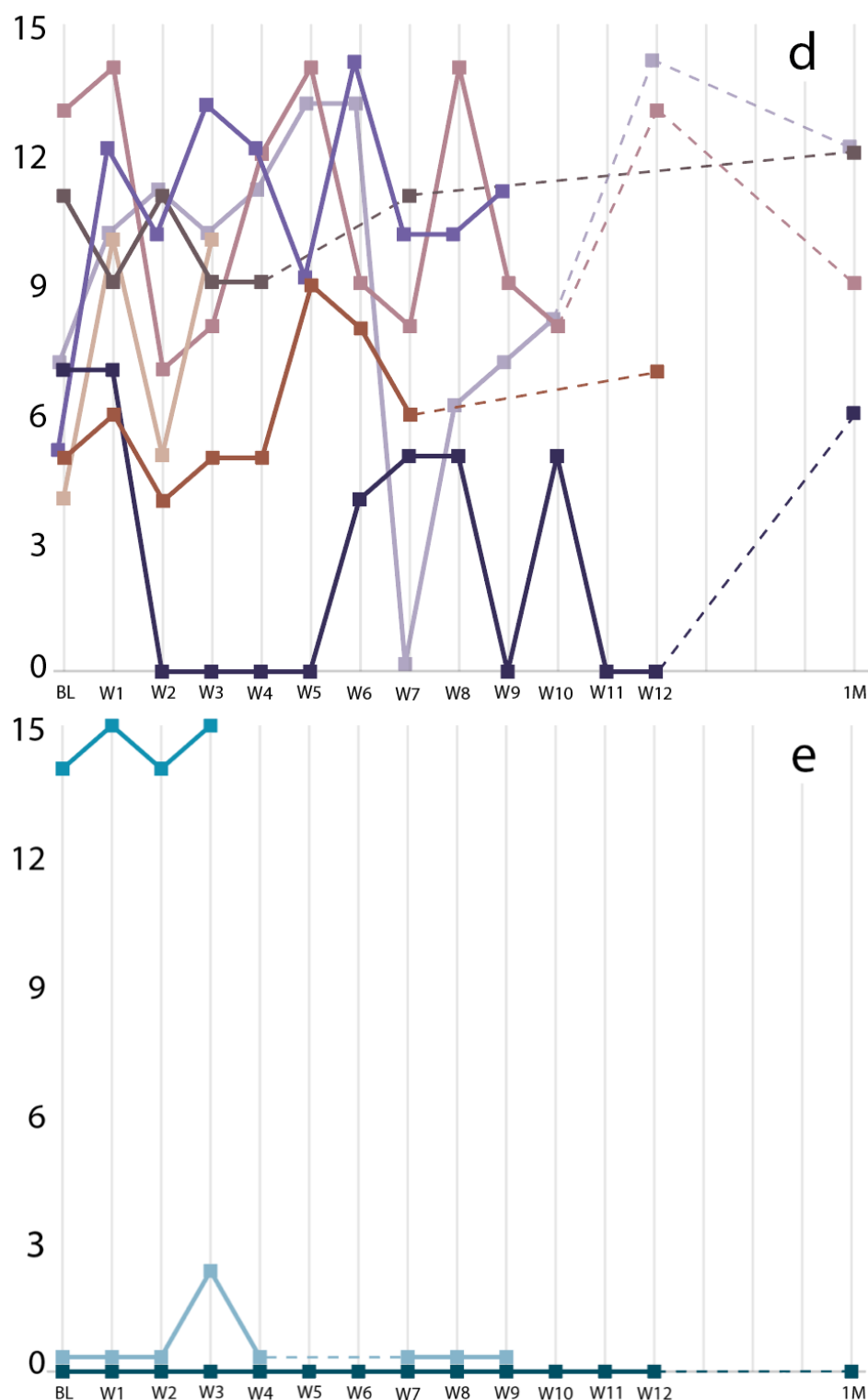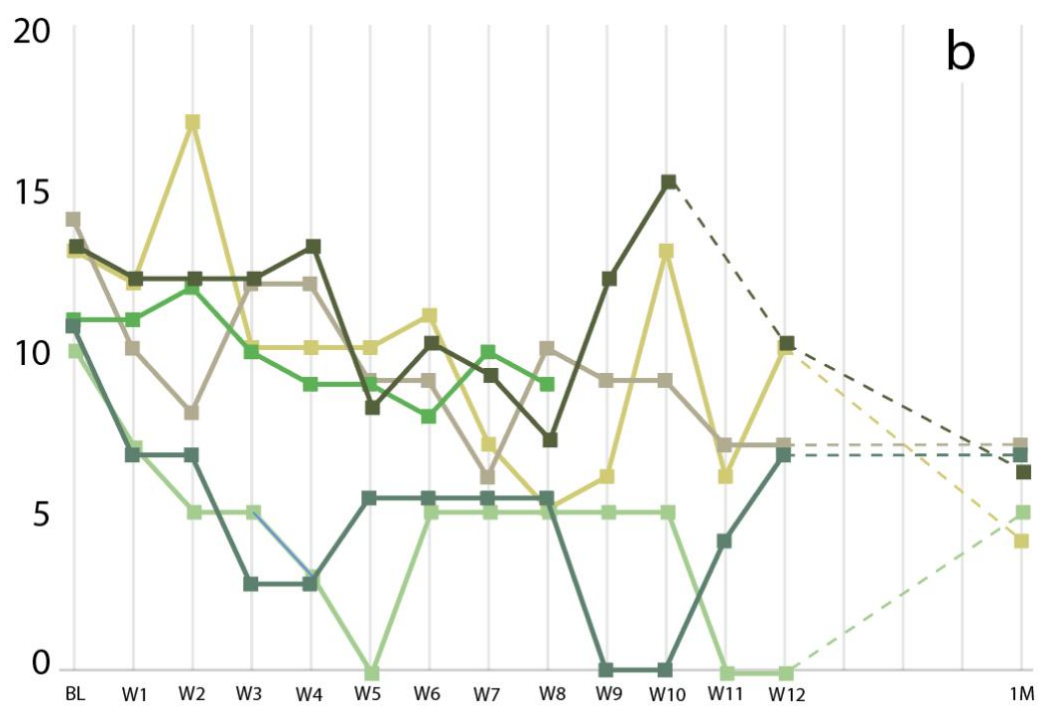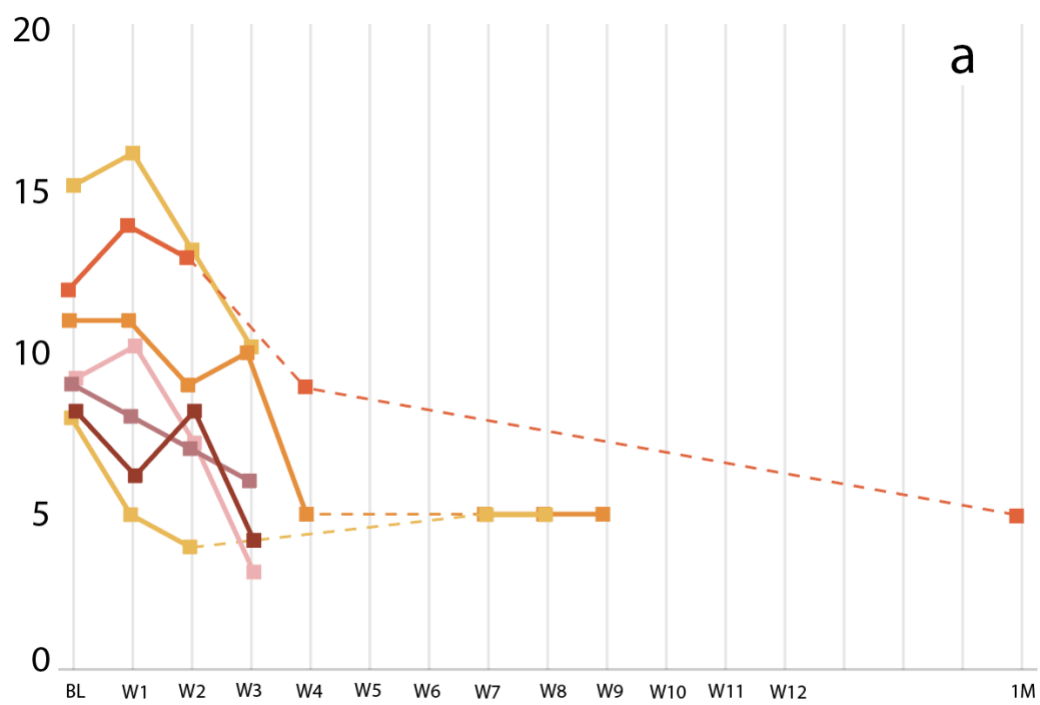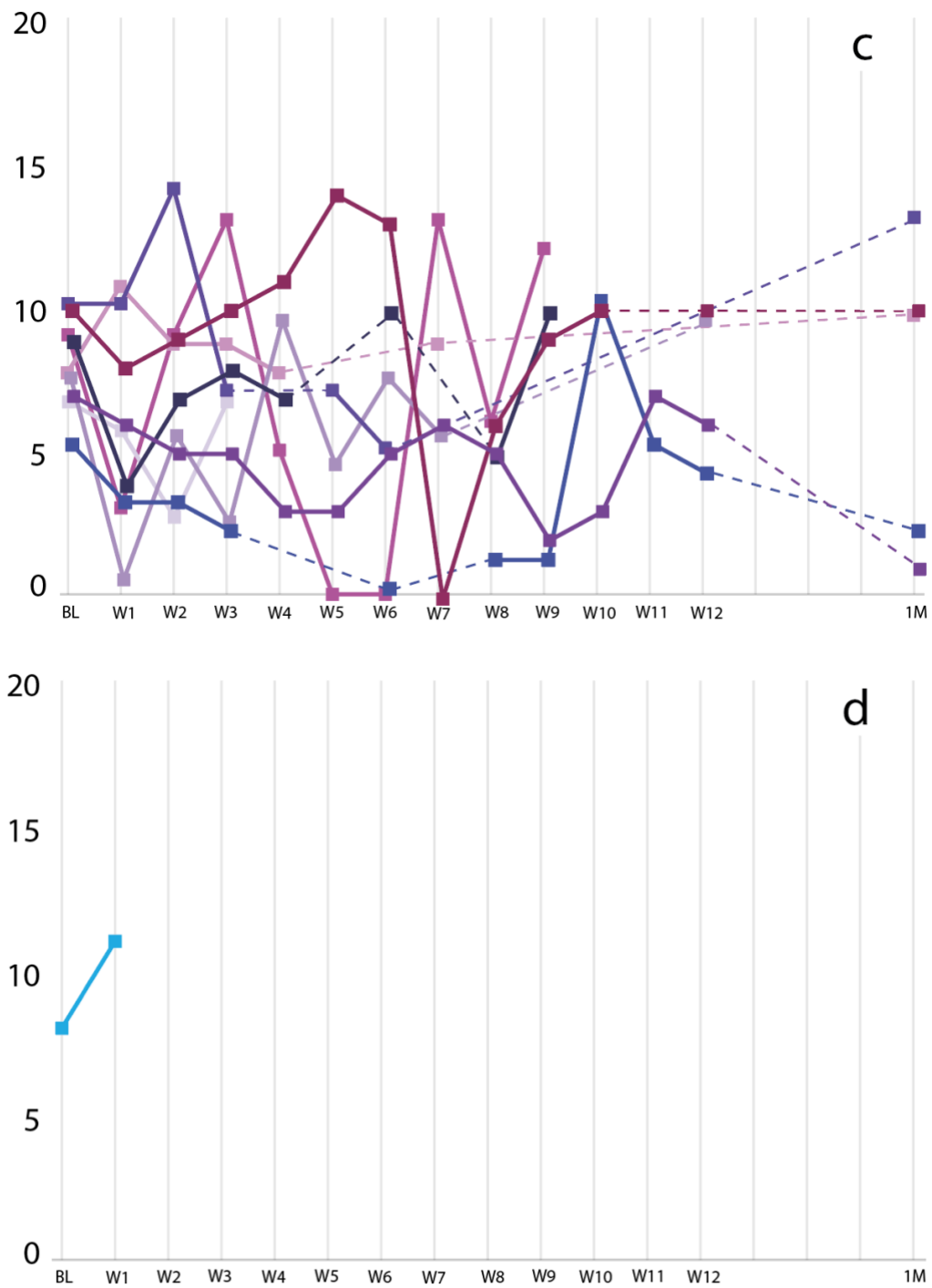
Figure 17. Individual OASIS trajectories, grouped by pattern. A. Consistent improvement, n=7. B. Improvement with cyclical pattern, n=6. C. Cyclical pattern with no improvement, n=9. Worsening symptoms, n=1.

Chapter 4. Discussion

This study offers one of the most comprehensive measurements of engagement outcomes in a digital health trial. We collected 10 total measures of engagement and assessed engagement at 14 different time points; the norm in the field is to collect one measure at one timepoint. To further deepen our understanding of engagement with the Hazel intervention, we utilized a combination of self-report, behavioral and objective measures of engagement. There was high acceptability of the Hazel intervention design and measurement schedule, and participant feedback about the intervention was overall positive. While there was little evidence that the response burden of our measures was too high for participants, we observed that some of our measures of engagement showed little variability between participants or over time. These results offer some of the first empirical guidance on the selection of measures of engagement and the frequency of their administration.

Surprisingly, we found that therapist support increased some aspects of intervention usage and decreased other aspects; this nuance would have been lost if we had used only one measure of technological engagement. In addition to underscoring the utility of more comprehensive measurement of engagement, these findings provide guidance on potential adaptations to the trial design that might improve engagement. Even with the current design, we found preliminary evidence that most participants had some reduction in symptoms over the course of the study, suggesting the Hazel intervention may be effective for improving problems related to depression and anxiety. Notably, this was despite minimal evidence of change on the underlying theoretical

mechanisms of the intervention or changes in the behaviors the intervention targets (i.e., cognitive flexibility, mindfulness, avoidance).

Below, we discuss the implications of our findings in greater detail.

4.1 Acceptability of the Hazel Intervention

Participants in both arms showed relatively high experiential engagement at all timepoints, suggesting high acceptability of the intervention. System Usability Scale scores were overall positive for both arms at both baseline and 12 weeks, and participants were generally more likely than not to say they would recommend Hazel to others.

There is likely little clinical significance to the 2.7-point difference in SUS scores between arms at baseline or the 3.4-point difference in SUS scores at 12 weeks between arms. Similarly, the 2.2-point decrease in mean SUS scores from baseline to 12 weeks for the supported arm is unlikely to represent a true decrease in attitudes toward the intervention. Notably, the standard deviation for this arm is elevated at 12 weeks and substantially higher than the standard deviation for all other arms and timepoints, suggesting greater heterogeneity in attitudes toward the intervention for participants in the supported arm upon completion of the intervention, as reflected in the scores ranging from 30 to 100. However, all interpretations are limited by the low number of respondents in this arm post-intervention.

In contrast, the 8.4-point decrease in mean SUS score for participants in the unsupported arm may be indicative of worsening attitudes toward the intervention over time, from "above average" attitudes at baseline to "average" attitudes after completing the intervention. With the currently available data, it is unclear if this trend is unique to the unsupported arm or if a similar trend would emerge if there were more participants in the supported arm. If this decrease represents a true decrease in favorability toward

the intervention, this might reflect intervention delivery decisions that were made for feasibility rather than participant experience. Future, fully powered studies may seek to better understand changes in SUS over time, as well as any impact that the presence/absence of human support has on this change.

Participants in the unsupported arm were more likely to express willingness to recommend the intervention to others. These differences were unexpected, as participants were blinded to their condition at baseline and there were no differences in the intervention until Week 8, when therapist support began. It is therefore unlikely that these baseline differences between arms are attributable to our experimental manipulation. All subsequent differences between arms should be interpreted with caution due to this baseline difference. Interpretation of these trends is further limited by the low number of respondents in later weeks of the intervention, especially in the supported arm. As noted above, participants in the unsupported arm were more likely to answer this item in later weeks, though this may be an unintended consequence of the study design. Notably, most participants who completed the video sessions offered positive verbal feedback about their experiences.

The overall positive feedback about the Hazel intervention is especially notable as the intervention website was designed primarily for feasibility rather than to optimize the user experience. The intervention was delivered over RedCap, a secure web application for data collection. While RedCap surveys are customizable, options are limited and the resulting Hazel intervention was relatively simple in appearance and functionality (see **Figure 18** for an example of the intervention website). Therefore, we anticipate that experiential engagement outcomes might be greater for studies using a more modern

design, so long as caution was taken to ensure the website was still intuitive for new

users.

Figure 18. Screenshot of the participant view of Week 1 of the Hazel intervention.

4.2 Feasibility of the Hazel Intervention

One of the primary aims of this exploratory study was to determine the feasibility

of our experimental manipulation. Specifically, we sought to determine if there was

sufficient retention and data availability to assess differences between arms when video

sessions started in Week 8 and to assess the therapist burden of four 50-minute sessions

per participant in the supported arm.

4.2.1 Timing of video sessions

Unexpectedly, participants in the supported arm showed worse technological engagement across numerous measures tied to the onset of video sessions. Specifically, participants in the supported arm were less likely to start weekly lessons when video sessions began, the majority (63.64%) completed no video sessions, and there was a substantial spike in non-use attrition during weeks 8 and 9 for these participants. Taken together, our weekly completeness and attrition outcomes indicate that not only were participants unlikely to complete video sessions, they were unlikely to continue in the intervention at all once video therapy sessions were offered. There are several possible explanations for this and corresponding adjustments that might be made to future studies.

Firstly, it is possible that our choice to deliver the video therapy sessions relatively late into the intervention, during weeks 8 through 11, may have contributed to lower rates of video session completion. Participants may have gotten familiar with completing the intervention independently and felt uncomfortable initiating human support after 7 weeks. Future studies might experiment with delivering support at different timepoints to observe how this changes engagement.

It is also possible that the introduction of an *unknown* therapist contributed to lower engagement. Notably, two of the participants with the highest number of video sessions completed (4 sessions and 2 sessions, respectively) were assigned to work with the study's primary investigator (JB). All participants were screened by the PI by phone, and it is therefore possible that this initial contact led to increased comfort beginning video therapy sessions for those assigned to the PI versus those assigned to other study

therapists. Future trials might also try introducing the assigned study therapist earlier on to see if this impacts engagement.

Lastly, we chose to keep participants blinded to their condition until Week 7 of the intervention, when those in the supported arm were contacted by the PI and their assigned study therapist to schedule video sessions. Our intention was to minimize differences in the intervention between arms before the onset of the video sessions. However, it is possible that if participants were alerted to their condition and were anticipating video sessions starting there might have been greater uptake once the video sessions began. Again, future studies might experiment with when participants are informed of their condition to see how changes in participant expectations impact engagement.[65]

4.2.2 Number of video sessions

In total, we conducted 9 50-minute sessions, for a total of 450 minutes or 7.5 hours over 9 months of data collection. This was substantially lower than we expected (4 50-minute sessions times 13 participants, for a total of 43.3 anticipated hours). In order to meet this expected burden, this trial had a team of five study therapists. Given the low uptake of video therapy sessions, it was highly feasible for a team of this size to offer this dose of support. However, if changes to the intervention were made to attempt to increase uptake of video sessions, the feasibility of offering 4 50-minute sessions per participant should continue to be monitored. As noted above, these changes might include reconsidering when video sessions are offered. We decided to offer 4 sessions primarily based on the length of the emotion exposure module. If video sessions were moved to the start of the intervention, there would be no rationale for offering 4 video sessions rather

than one, two, or three. Reducing the number of video sessions is one option to reduce the negative impacts of increased video session uptake on therapist burden and intervention scalability.

4.2.3 Participant Retention

Overall, fewer participants started each weekly lesson as the intervention went on. For those who did start each lesson, the level of completeness also decreased over the course of the intervention. This aligned with our expectations and the norms for digital health interventions. Homework completion generally decreased over the course of the intervention, as we expected based on findings from other digital health studies.

4.2.3.1 Weekly completeness

Notably, for the minority of participants in the supported arm who did complete video sessions, there was a higher average level of completion during weeks 8-11 than for participants in the unsupported arm. This aligned with our expectations that the in-vivo support and feedback of a therapist would improve engagement with intervention content during this challenging module. However, this boost in lesson completeness is substantially hampered by the much lower number of participants in the supported arm who started any video lessons. In sum, our findings indicate that offering therapist support, as delivered in the current study, was beneficial for those who utilized it but overall discouraged participants from continuing in the intervention.

One notable limitation of *completeness* as an outcome is that it does not account for actual understanding of the concepts being taught, how much effort was put into responses, or even if responses are relevant in any way. Completeness was simply calculated as the number of responses a participant gave in a given week divided by the

total number of responses possible; the content of those responses was not included in any way. Notably, there is some evidence that participants did not understand all topics or accurately complete all lessons. For example, one participant would enter the word "Unknown" when confused about a topic. Another answered a prompt to identify jumping to conclusions and catastrophic thinking by noting that they "never" engage in either; however, there were numerous examples of both in their responses to other prompts within the same lesson. For both of these participants, completeness would indicate higher levels of engagement than a content-based measure. This is a major caveat of this measure, as simply completing an intervention without accurately understanding the concepts is unlikely to lead to benefits.

4.2.3.2 Homework completeness

Notably, both the length (i.e., how many total responses participants were prompted to give) and frequency (i.e., how many days participants were asked to practice the skill) of homework assignments varied by week. It appears that both frequency and length of assigned homework impacted completion. The two weeks with the highest average homework completion (weeks 1 and 6) were the only two weeks with one-time homework assignments; every other week participants were asked to practice the skill multiple times each week. Week 1, which had the highest homework completion on average, also had the shortest total homework assignment. However, several weeks had shorter assignments than Week 6, so it appears that homework *frequency* had a greater impact on completion than homework *length*. For homework assignments with a frequency of more than one day, the number of days did not appear to impact homework completion; rather, the total length of the homework appeared most impactful. Week 3

had the third-highest mean completion after the two weeks with one-time homework. Week 3 had daily practices that were the briefest after Week 1, while other weeks (specifically, weeks 9-11) had longer practices but only 4 times a week.

However, it is also possible that this pattern may in part be an artifact of our study design rather than accurately reflecting homework completion. It is possible that participants completed daily homework practices but failed to return to the study website to report on the practice. Participants were not incentivized to complete homework tracking (i.e., their financial compensation was not dependent on homework completion) and, as noted above, most participants used the intervention for only a single day per week. Participants would learn what the homework assignment was on the first day they used the intervention; if they never returned to that week's lesson it would be impossible for them to report their homework completion on the study website. Future iterations of the Hazel intervention might include prompts or incentives to complete homework tracking or change how homework completion was tracked (for example, shifting reporting about the homework to the start of following week's lesson rather than the end of the same week's lesson; asking briefer self-report questions about homework completion such as "How many days last week did you complete the home practice?").

Additionally, homework completion has similar limits as the weekly completion outcome discussed above: Completion does not account for the content of the homework completion, such as depth of completion or relevance of the data entered. There may be participants with a high level of completeness who have not understood or accurately completed the home practice, and vice versa. Future studies might consider alternatives to completeness, such as qualitative analysis of the *content* of home practice responses, to

better understand the relationship between intervention design, behavioral engagement, and other engagement and health outcomes.

4.2.3.3 Non-use attrition

The major limitation of non-use attrition as an outcome is that it fails to distinguish between reasons for non-use. There are numerous reasons why someone might stop using an intervention, and those reasons have different implications for the intervention itself. For example, someone might stop using an intervention because they've forgotten about it (a negative type of attrition), because they don't like the intervention (another negative type of attrition), or because their symptoms have been effectively reduced and they no longer feel that they need the intervention (a positive type of attrition, which might be better labeled as early intervention success).[24,25] Analyses contrasting attrition point with other outcomes are required to better understand the reason for attrition. For example, we could infer whether attrition was due to disliking the intervention by investigating the relation between attrition point and experiential engagement outcomes. Similarly, we could infer whether attrition was due to intervention success by investigating the relation between attrition point and symptom outcomes. The low sample size and high missingness of the current study preclude such analyses, but future studies may want to ensure they are sufficiently powered to test these relations between outcomes. Alternatively, future studies might seek to conduct exit interviews with participants to get self-reported reasons for non-use. All options have pros and cons, and more comprehensive approaches involving multiple of these strategies may be especially useful for future studies seeking to better understand not only patterns of non-use attrition but also the underlying reasons.

75

4.2.4 Intervention Effects

As an exploratory study with a small sample size, we cannot draw strong conclusions about the effectiveness of the Hazel intervention for reducing depression and anxiety symptoms, nor are we powered to do so. Rather, we aimed to establish whether there was any evidence of a treatment effect on depression symptoms, anxiety symptoms, or the underlying mechanisms that would support conducting future trials of the intervention.

4.2.4.1 Intervention Mechanisms

Overall, we observed minimal changes on our measures of the theoretical mechanisms of the intervention. At baseline, participants showed relatively high abilities to generate alternative perspectives on difficult situations and had more moderate confidence in their ability to respond effectively to difficult situations. The less-than one-point increase participants showed on the control subscale is very unlikely to be clinically significant; however, it is possible that the roughly 5-point decrease on the alternatives subscale does reflect a meaningful decrease in participants' ability to generate multiple interpretations and possible solutions for difficult situations. This is contrary to our hypotheses, as cognitive flexibility is one of the theoretical mechanisms of the Unified Protocol and two weekly lessons (weeks 4 and 5) are devoted to improving these skills.

Both baseline and post-intervention scores are indicative of moderate mindfulness. The approximately one-point increase from baseline to 12 weeks is very unlikely to represent a clinically significant increase in mindfulness. Again, this was contrary to our hypotheses as mindfulness is one of the major theoretical mechanisms of the Unified Protocol and mindfulness is incorporated into nearly all weekly lessons.

Similarly, it is unlikely that there is any clinical significance to the approximately 1-point decrease in score from baseline to post-intervention; both scores are indicative of moderate levels of avoidance. Again, this was contrary to our hypotheses, as targeting avoidance is another core theoretical mechanism of the Unified Protocol and a major focus of weeks 6 through 11.

In sum, these findings suggest that there may be issues in terms of how the Hazel intervention is written. Future studies of the Hazel intervention might review and revise the content of these lessons and measures. The Efficiency Model of Support notes that absence of change in behavioral, digital health interventions can be attributed to several causes.[44] If our intervention were failing to adequately teach the skills of cognitive flexibility, mindfulness, or approaching instead of avoiding, this would be considered a knowledge failure under the Efficiency Model of Support. However, if participants were accurately understanding the skills but failing to incorporate them into their daily lives (e.g., lack of behavioral engagement), the Efficiency Model would term this an implementation failure.

It is also possible that the lack of observed change on these outcomes was due to issues with how we measured or analyzed these outcomes, rather than true lack of change. Notably, despite minimal changes on all of the theoretical mechanisms of our intervention, most participants showed at least some reduction in depression and anxiety symptoms, though there was significant variability between participants. We expect that there should be a link between changes on these mechanisms and ultimate symptom change; it is possible that simply computing averages for these outcomes collapsed those who did experience improvements in these mechanisms with those who experienced no

or maladaptive changes. Future, fully powered studies might conduct multivariate analyses investigating the relation between these behavioral engagement outcomes and symptom change to better understand patterns in these outcomes.

4.2.4.2 Depression Symptoms

While average depression symptoms across all participants showed a decrease over time, interpretations of these averages are limited by the substantial variability in symptom trajectories between participants. When looking at individual trends, most participants showed some improvement in depression symptoms (n=13), and a sizable minority showed cyclical spikes and drops in symptoms with no substantial change over the course of the intervention (n=7). However, interpretation of individual trends is also limited by substantial missingness. For example, of the 8 participants who showed consistent reduction in the depression symptoms, 4 submitted data at four timepoints or fewer. While the data we have shows symptom reduction, it is possible that participants' symptoms increased at later timepoints for which we do not have data.

Similarly, there was substantial missingness for both participants who showed initial reduction in depression symptoms followed by a spike at the last available datapoint, making it impossible for us to conclude whether these upticks represent one-off spikes, cyclical patterns, or true symptom non-improvement. Notably, two of the three participants who showed no change in depression symptoms had minimal symptoms at all timepoints, so the absence of symptom change does not have implications for the effectiveness of the intervention. The third who showed no improvement had clinically significant symptoms but completed only three weeks of the intervention, again limiting our ability to draw conclusions about the reason for non-recovery.

78

Of note, there was much less missingness among participants who showed cyclical patterns in their symptoms, including both those who showed overall reduction in their symptoms over time and those who showed no change in symptoms over time. We can therefore feel more confident in interpreting these results.

Above, we discussed possible changes to the trial design to increase technological engagement. If these changes were effective, they would likely also increase completion of symptom monitoring surveys as these surveys were the first thing participants completed in each weekly lesson. Future studies might also incentivize completion of symptom surveys separate from completion of weekly lessons to increase availability of these data. Resolving issues of symptom data missingness is particularly important for any studies interested in the effectiveness of the intervention, as effectiveness cannot be established with the current levels of completion.

4.2.4.3 Anxiety Symptoms

Anxiety symptoms had similar trajectories as depression symptoms, and their interpretation is similarly limited by missingness. Average anxiety symptoms across all participants decreased over time, with a notable increase in weeks 9 and 10. This increase may be attributable to the onset of exposure-based activities in weeks 9-11; we would expect anxiety symptoms to rise at the onset of exposure and to decrease with continued exposure. Notably, this spike becomes more pronounced when non-anxious participants are removed, consistent with this interpretation. However, the low number of participants in these weeks and the significant variability in individual trajectories limits our ability to conclusively interpret these patterns.

When looking at individual trends, most participants showed some improvement in anxiety symptoms (n=13) over the course of the intervention, and a substantial minority showed cyclical spikes and drops in symptoms with no substantial change over the course of the intervention (n=9). However, interpretation of individual trends is also limited by substantial missingness. Of the 7 participants who showed consistent reduction in anxiety symptoms, 3 submitted data at four time points or fewer. While the data we have shows symptom reduction, it is possible that participants' symptoms increased at later timepoints for which we do not have data. Similarly, the one participant who showed worsening anxiety symptoms completed only one week of the intervention. As with symptom improvement, this missingness prevents us from drawing firm conclusions about the effectiveness of the intervention. Again, there was less missingness among those with cyclical patterns to their symptoms, improving our ability to interpret these results with confidence.

It is likely that the changes mentioned above to increase technological engagement and depression symptom monitoring would also improve anxiety symptom monitoring.

4.3 Feasibility of the Trial Design

This trial aimed to establish the feasibility of collecting numerous measures of engagement at greater frequency than is typically done. We found no evidence of response burden for our measures, but we did find that some provided less useful information than others. We also attempted to recruit a highly diverse sample; our sample was more diverse than previous studies and the general US population in terms of

LGBTQ+ identity but not racial/ethnic identity or socioeconomic status. Below, we discuss takeaways from the results of this trial and possible changes for future studies.

4.3.1 Measures

Most of our measures of engagement provided useful information for better understanding the Hazel intervention. The System Usability Scale, willingness to refer, weekly lesson completion, non-use attrition, weekly homework completion, Cognitive Flexibility Index, Cognitive and Affective Mindfulness Scale, and COPE Inventory all produced sufficient data with sufficient variability to help us more comprehensively understand engagement with Hazel. However, the number of referrals made and the number of days per week that participants used the intervention offered relatively little variability, indicating they may not contribute much to our understanding of engagement with this intervention.

4.3.1.1 Number of Referrals

Participants made very few actual referrals. We are not aware of any previous trials that have used number of referrals as an engagement outcome. This strategy is more often seen in snowball sampling; previous research has indicated that a variety of factors can impact the success of snowball sampling methods.[66] Most notably for this study, this can include when the participant feels the research topic is highly personal and therefore does not want to disclose information about their own participation to others and when there are weak social connections between participants and potential referrals. Disclosure of participation in a mental health study may have been too personal for many of our participants to feel comfort

81

making referrals; it may also have been considered too personal to suggest to someone else that they might need treatment for mental health symptoms. Further, depression and anxiety often lead to isolation and challenges with maintaining social relationships, which may have further limited the feasibility of participant referrals.

We included this as a novel variable in an attempt to move beyond simply self-report measures of experiential engagement, but our findings indicate that actual referrals may not be the best behavioral measure of experiential engagement. Future studies may consider alternative measures of experiential engagement, including both self-report (e.g., open-ended questions about participants' experiences of the intervention; exit interviews) and behavioral measures (e.g., eyetracking; observed usability testing).

4.3.1.2 Days used per week

There was very little variability between participants in the number of days per week that they used the intervention; most consistently used the intervention only one day per week. While this does help us better understand discrepancies between intervention design (e.g., for several weeks participants were encouraged to track their home practice daily) and participants' use of the intervention (e.g., at no point did any participant track their home practice daily), this may not be the most informative technological engagement outcome for a full RCT. Rather, this engagement outcome may be most helpful in pilot studies to help interventionists better understand mismatches between the way they intended for an intervention to be used and how participants are actually using it. At the same time, because this outcome can be tracked automatically by the intervention website, it does not contribute to response burden and there is therefore little harm in collecting it.

Notably, all of the technological engagement outcomes for this trial were behavioral; future studies might explore non-behavioral measures such as self-report questionnaires about the effort and attention participants put into the weekly lesson. Alternatively, the content of participants' responses might be coded to attempt to behaviorally capture the attention/effort participants have put into using the intervention. Of course, non-comprehension might also indicate a problem with the intervention itself rather than a lack of effort on the part of participants. Coding qualitative data might help identify any parts of the intervention itself that would benefit from being revised, as we would expect drops in comprehension/accuracy outcomes across most participants at these points.

4.3.2 Assessment schedule

While the SUS and willingness to refer provided useful information, results of this trial suggest possible changes to the frequency with which these measures were administered.

4.3.2.1 Days used per week

Notably, numerous participants reached out to the primary investigator with questions about the SUS when completing it at baseline. Though participants had used the Hazel website when completing the online screener, consent form and baseline questionnaires, they seemed unaware that this was the same platform as the Hazel intervention and were therefore uncertain of how to answer the SUS questions. Thus, it may make more sense to administer the SUS immediately following Week 1 of the intervention for future studies.

4.3.2.2 Willingness to recommend

As an engagement outcome, willingness to recommend an intervention to others is most often collected once following completion of the intervention; however, there was little evidence to support this administration schedule. If any rationale were given, it typically focused on response burden. As noted above, we found little evidence of response burden in our sample despite the large number and frequency of assessment of engagement. We hypothesized that willingness to refer might vary weekly based as participants gained additional experience with the intervention technology and content, but results indicated that participants were generally consistent in their answers, rather than showing much change week to week. This indicates that this willingness to refer might not need to be collected weekly; rather, it might be collected at a few points over the course of the intervention to lessen response burden. Willingness to recommend an intervention is most commonly collected only at post-intervention.

Notably, there were differential response rates in willingness to refer between arms, with participants in the supported arm less likely to respond to this item in the weeks with therapist support compared to those in the unsupported arm. While these differences are largely driven by disparities in lesson completion between the two arms during these weeks, they may also reflect how content of each weekly lesson was tunneled. Participants had to complete and submit weekly symptom surveys in order to see the lesson containing the weekly psychoeducational reading, practice exercises and home practice assignment. They had to submit this lesson in order to see the item asking if they would recommend Hazel to others. However, most participants who completed video sessions did not fill out and submit the weekly lesson on the intervention website;

84

rather, they completed the lesson verbally with their therapist and reported their homework verbally at the start of the next session. Thus, because these participants did not submit the weekly lesson, they were not shown the item asking about their willingness to refer the intervention to others on weeks when they completed video sessions. Future studies including weekly assessment of outcome measures should be cognizant of how content tunneling and intervention design may impact data collection.

4.3.3 Participants and recruitment

Compared to many other digital health studies, our sample was more diverse in terms of LGBTQ+ identity (21%) and roughly comparable to other studies in terms of educational attainment (40% with an associate's degree or less) and BIPOC-identified participants (17%). Across the 42 studies included in Brooks et al.'s review of the impact of study design on experiential engagement, relatively few reported on participant demographics.[15] Of those that did, 75% of participants were white and 60% had a 4-year degree or higher. None of the trials reported on LGBQ+-identified participants and none reported any trans- or gender-diverse participants. Notably, our sample was also less racially/ethnically diverse than the US population (roughly 76% white per the 2021 census) or the state of North Carolina (70% white).[67] Estimates of the percentage of the US population that identifies at LGBTQ+ have been rapidly changing in recent years; a 2021 Gallup survey found that 7.1% of adults in the US identified as LGBTQ+.[68]

Overall, this suggests that our recruitment strategies were effective for oversampling LGBTQ+-identified individuals but not BIPOC or low-SES participants. While we attempted a multi-pronged recruitment strategy including multiple sources that primarily serve BIPOC, low-SES and LGBTQ+ individuals, we had little success with

these recruitment sources. Rather, most of our sample was recruited from

ResearchMatch, which reports that 75% of the registry population is white and 0.6% is

transgender; other demographics are not reported.[69] It is also important to note that

participants from ResearchMatch may not be representative of the US population in terms

of technological literacy, which is a particularly relevant concern for digital health

studies. Future studies should prioritize exploring and identifying effective strategies for

diverse recruitment of BIPOC and low-SES populations.

Further, we operationalized SES exclusively through educational attainment, but

this is only one aspect of socioeconomic status. As there was limited evidence that

participants found the response to current measures too burdensome, future studies might

incorporate additional metrics of SES such as income, occupation, employment status,

citizenship status, and/or questions about class-related stressors (e.g., housing or food

insecurity, ability to access necessary medical care, access to sick leave, et cetera).

4.3.4 Compensation structure

There was a spike in non-use attrition between weeks 2 and 3 for both arms of our

trial. Notably, participant compensation was dependent on intervention completion, with

participants who completed at least 2 weeks of the intervention receiving $60. While this

compensation structure was intended to incentivize building the habit of using the

intervention and ensure at least 2 time points of data for analyses, it is possible that it

unintentionally contributed to this observed pattern in attrition. Further, there are

significant limitations to the conclusions we can draw based on only two weeks' worth of

data. A more gradual compensation structure or a greater number of sessions completed

to earn compensation might lead to reductions early intervention attrition.

4.4 Strengths and Limitations

This study has made several novel contributions to the field of digital health research. Most notably, we conducted much more comprehensive assessment of engagement outcomes than is typically seen. Our results provide some of the first empirical guidance on the selection of measures of engagement and the frequency with which these measures should be administered. At the same time, "engagement" is a vast construct, and there are still many potentially important aspects of engagement that we did not assess in this study.

Another novel contribution of this study was the focus on a comparatively low dose of support. The Unified Protocol, upon which the Hazel intervention is based, has thus far only been tested in trials with weekly therapist support.[49,70–73] More broadly in the digital health realm, most research on the impact of therapist support has contrasted a standalone arm with an arm that receives support throughout the intervention period.[15,30–42,74] Standalone interventions are much more scalable that supported ones, which is an important consideration when considering the promise of digital health. While our design is much more scalable than a comparable intervention with weekly therapist support, we still planned to deliver 200 minutes of therapist support per participant in the supported arm. This is much less scalable than the typical standalone digital health intervention. Arguments have even been made in favor of scalability over effectiveness — from a public health standpoint, there is more benefit from an intervention that only works for 1% of users but can serve 100,000 (1,000 individuals successfully treated) than an intervention that works for 80% of users but can only serve 100 (80 individuals successfully treated).[75] The extent to which support impacts outcomes will likely vary

based on the condition being treated and the target population. Ultimately, we believe it will be most beneficial for interventionists to make decisions about the dose of support based on both scalability and effectiveness for their specific situation.

This study was also unique in its efforts to recruit a highly diverse sample for a trial of a digital mental health intervention. We successfully oversampled participants who identified as members of the LGBTQ+ community, which has historically been underrepresented and underreported in research. However, we were unsuccessful in our attempts to oversample participants from underrepresented racial/ethnic backgrounds and lower-SES participants. One major limitation of the current study is that nearly all participants were recruited from ResearchMatch and may not be representative of the U.S. population.

There are both benefits and drawbacks of our exploratory design. Most notably, our low sample size means we are underpowered to explore the statistical significance of our findings or to draw firm conclusions about the effect of our intervention on engagement or health outcomes. Our ability to draw conclusions is further limited by substantial missingness. At the same time, this study was never designed to be fully powered. Rather, it was intended as a small-scale feasibility trial of our design to identify issues before conducting a fully powered trial, and it was successful in accomplishing those aims.

Further, we hope this study has made meaningful contributions to the field of pilot study methodology. Over the past several years, numerous arguments have been made against the use of inferential statistics for evaluating the results of pilot trials. However, there is little consensus about how best to make use of and interpret pilot data. In addition

88

to descriptive statistics, we made extensive use of data visualizations to help us better understand trends across, between and within participants over the course of the intervention. We believe these visualizations substantially deepened our understanding of the feasibility and acceptability of the current trial design and that similar methods could be helpful in evaluating other pilot studies.

4.5 Conclusions

This exploratory trial successfully developed and delivered an online, mental health intervention to 23 participants. We randomized participants to complete identical interventions based on the empirically supported Unified Protocol, with only one tightly controlled difference between arms. This contributes to recent efforts to open the "black box" of intervention effects and better understand how each individual component contributes to outcomes. We successfully accomplished our aims of determining the acceptability and feasibility of the current intervention and trial design, and we identified several possible changes that might increase participant retention and data availability for future trials. Even without these changes, the Hazel intervention showed high acceptability among participants and showed preliminary evidence of efficacy for reducing depression and anxiety symptoms.

Appendix A: Behavior Change Techniques in this intervention

We identified the Behavior Change Techniques present in the Unified Protocol by coding the Unified Protocol manual.[27] Coding was performed by the primary investigator (JB), who completed online training in identifying and coding BCTs.[39] Identified BCTs were then confirmed in consultation with one of the authors of the manual (Clair Cassiello-Robbins).

Table 9. Behavior Change Techniques in the Hazel intervention.

No.	Name	No.	Name
	Core Components of the Unified Protocol		
1.1	Goal Setting (behavior)	5.4	Monitoring of Emotional Consequences
1.2	Problem Solving	5.6	Information about Emotional Consequences
1.3	Goal Setting (outcome)	6.1	Demonstration of the Behavior
1.4	Action Planning	7.7	Exposure
1.5	Review Behavioral Goal(s)	8.1	Behavioral Practice/Rehearsal
1.6	Discrepancy Between Current Behavior and Goal	8.2	Behavioral Substitution
1.7	Review Outcome Goal(s)	8.3	Habit Formation
2.2	Feedback on Behavior	8.4	Habit Reversal
2.3	Self-Monitoring of Behavior	8.6	Generalization of Target Behavior
2.4	Self-Monitoring Outcome(s) of Behavior	8.7	Graded Tasks
2.7	Feedback on Outcome(s) of Behavior	9.2	Pros and Cons

4.2	Information about Antecedents	11.4	Paradoxical Instructions
4.3	Re-Attribution	13.2	Framing/Reframing
4.4	Behavioral Experiments	13.3	Incompatible Beliefs
Additional Components Included in this Intervention			
3.2	Social support (practical)	3.3	Social support (emotional)
7.1	Prompts/Cues	10.2	Material Reward (Behavior)
10.1	Material Incentive (Behavior)	10.4	Social Reward

References

1.   Neary M, Schueller SM. State of the Field of Mental Health Apps. *Cogn Behav Pract*. Published online March 5, 2018. doi:10.1016/J.CBPRA.2018.01.002

2.   eMarketer. *Smartphone Penetration in the US 2010-2021*.; 2016. Accessed December 13, 2016. https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

3.   eMarketer. *United States Mobile Phone Penetration 2010-2018*.; 2016. Accessed December 13, 2016. https://www.statista.com/statistics/222307/forecast-of-mobile-phone-penetration-in-the-us/

4.   eMarketer. *Global Smartphone User Penetration 2014-2020*.; 2016. Accessed February 14, 2017. https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/

5.   eMarketer. *Mobile Phone Penetration Worldwide 2013-2019*.; 2016. Accessed February 14, 2017. https://www.statista.com/statistics/470018/mobile-phone-user-penetration-worldwide/

6.   Apolinário-Hagen J, Kemper J, Stürmer C. Public Acceptability of E-Mental Health Treatment Services for Psychological Problems: A Scoping Review. *JMIR Ment Heal*. 2017;4(2):e10. doi:10.2196/mental.6186

7.   Zhao J, Freeman B, Li M. Can Mobile Phone Apps Influence People's Health Behavior Change? An Evidence Review. *J Med Internet Res*. 2016;18(11):287-2871. doi:10.2196/jmir.5692

8.   Richards D, Richardson T. Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clin Psychol Rev*. 2012;32(4):329-342. doi:10.1016/j.cpr.2012.02.004

9.   Huguet A, Rao S, McGrath PJ, et al. A Systematic Review of Cognitive Behavioral Therapy and Behavioral Activation Apps for Depression. Choo K-KR, ed. *PLoS One*. 2016;11(5):e0154248. doi:10.1371/journal.pone.0154248

10.   Zabrinski MF, Celio AA, Jacobs MJ, Manwaring J, Wilfley DE. Internet-based prevention of eating disorders. *Eur Eat Disord Rev*. 2003;11(3):183-197. doi:10.1002/erv.525

11.   Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*. 2017;16(3):287-298. doi:10.1002/wps.20472

12.   Fjeldsoe BS, Marshall AL, Miller YD. Behavior Change Interventions Delivered

by Mobile Telephone Short-Message Service. *AMEPRE*. 2009;36(2):165-173. doi:10.1016/j.amepre.2008.09.040

13. Ruwaard J, Broeksteeg J, Schrieken B, Emmelkamp P, Lange A. Web-based therapist-assisted cognitive behavioral treatment of panic symptoms: A randomized controlled trial with a three-year follow-up. *J Anxiety Disord*. 24:387-396. doi:10.1016/j.janxdis.2010.01.010

14. Webb TL, Joseph J, Yardley L, Michie S. Using the Internet to Promote Health Behavior Change: A Systematic Review and Meta-analysis of the Impact of Theoretical Basis, Use of Behavior Change Techniques, and Mode of Delivery on Efficacy. *J Med Internet Res*. 2010;12(1):e4. doi:10.2196/jmir.1376

15. Brooks TL. A scoping review of the impact of intervention design on experiential engagement.

16. MacApagal K, Bhatia R, Greene GJ. Differences in Healthcare Access, Use, and Experiences Within a Community Sample of Racially Diverse Lesbian, Gay, Bisexual, Transgender, and Questioning Emerging Adults. *LGBT Heal*. 2016;3(6):434-442. doi:10.1089/lgbt.2015.0124

17. Glick JL, Theall KP, Andrinopoulos KM, Kendall C. The role of discrimination in care postponement among trans-feminine individuals in the U.S. national transgender discrimination survey. *LGBT Heal*. 2018;5(3):171-179. doi:10.1089/lgbt.2017.0093

18. Lykens JE, Leblanc AJ, Bockting WO. Healthcare experiences among young adults who identify as genderqueer or nonbinary. *LGBT Heal*. 2018;5(3):191-196. doi:10.1089/lgbt.2017.0215

19. Statista. *Mobile App User Retention Rate 2018*.; 2020. Accessed September 25, 2020. https://www.statista.com/statistics/751532/worldwide-application-user-retention-rate/

20. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Med Inform Decis Mak*. 2016;16(1):120. doi:10.1186/s12911-016-0359-3

21. Ricciardi L, Mostashari F, Murphy J, Daniel JG, Siminerio EP. A National Action Plan To Support Consumer Engagement Via E-Health. Published online 2019. doi:10.1377/hlthaff.2012.1216

22. Dentzer S. Rx For The 'Blockbuster Drug' Of Patient Engagement. *Health Aff*. 2013;32(2):202-202. doi:10.1377/hlthaff.2013.0037

23. Lupton D. The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Soc Theory Heal*. 2013;11(3):256-270. doi:10.1057/sth.2013.10

24. Yardley L, Spring BJ, Riper H, et al. Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *Am J Prev Med*. 2016;51(5):833-842. doi:10.1016/J.AMEPRE.2016.06.015

25. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Transl Behav Med*. 2017;7(2):254-267. doi:10.1007/s13142-016-0453-1

26. Prochaska JO, Velicer WF. The Transtheoretical Model of Health Behavior Change. *Am J Heal Promot*. 1997;12(1):38-48. doi:10.4278/0890-1171-12.1.38

27. HOWARD LEVENTHAL, IAN BRISSETTE. The common-sense model of self-regulation of health and illness. Published online August 21, 2012:56-79. doi:10.4324/9780203553220-10

28. Michie S, Richardson M, Johnston M, et al. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Ann Behav Med*. 2013;46(1):81-95. doi:10.1007/s12160-013-9486-6

29. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. *Heal Psychol*. 2008;27(3):379-387. doi:10.1037/0278-6133.27.3.379

30. Beiwinkel T, Eißing T, Sociology MA, Telle N, Siegmund-schultze E. Effectiveness of a Web-Based Intervention in Reducing Depression and Sickness Absence : Randomized Controlled Trial Corresponding Author : 19:1-14. doi:10.2196/jmir.6546

31. Berger T, Caspar F, Richardson R, Kneubühler B, Sutter D, Andersson G. Behaviour Research and Therapy Internet-based treatment of social phobia : A randomized controlled trial comparing unguided with two types of guided self-help. *Behav Res Ther*. 2011;49(3):158-169. doi:10.1016/j.brat.2010.12.007

32. Fitzpatrick KK, Darcy A, Vierhile M, Darcy A. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial Corresponding Author : 2017;4:1-11. doi:10.2196/mental.7785

33. Johnston N, Bodegard J, Jerström S, Åkesson J. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients : A randomized study. *Am Heart J*. 178:85-94.

doi:10.1016/j.ahj.2016.05.005

34. Morris RR. Efficacy of a Web-Based , Crowdsourced Peer-To-Peer Cognitive Reappraisal Platform for Depression : Randomized Controlled Trial Corresponding Author : 2015;17. doi:10.2196/jmir.4167

35. Mcclure JB, Anderson ML, Bradley K, An LC, Sheryl L. Evaluating an Adaptive and Interactive mHealth Smoking Cessation and Medication Adherence Program : A Randomized Pilot Feasibility Study Corresponding Author : 2016;4. doi:10.2196/mhealth.6002

36. Ho FY, Chung K, Yeung W, Ng TH, Cheng SK. Behaviour Research and Therapy Weekly brief phone support in self-help cognitive behavioral therapy for insomnia disorder : Relevance to adherence and ef fi cacy. *Behav Res Ther*. 2014;63:147-156. doi:10.1016/j.brat.2014.10.002

37. Alley S, Jennings C, Plotnikoff RC, Vandelanotte C, Alley S. Web-Based Video-Coaching to Assist an Automated Computer-Tailored Physical Activity Intervention for Inactive Adults : A Randomized Controlled Trial Corresponding Author : 18:1-15. doi:10.2196/jmir.5664

38. Mira A, Bretón-lópez J, Quero S, Baños RM, Botella C. An Internet-based program for depressive symptoms using human and automated support : a randomized controlled trial. Published online 2017:987-1006.

39. Unick AJL, Dorfman L, Leahey TM, Wing RR. A preliminary investigation into whether early intervention can improve weight loss among those initially non-responsive to an internet-based behavioral program. Published online 2016:2-7.

40. Hales S, Turner-mcgrievy GM, Wilcox S, et al. International Journal of Medical Informatics Social networks for improving healthy weight loss behaviors for overweight and obese adults : A randomized clinical trial of the social pounds off digitally (Social POD) mobile app ☆. *Int J Med Inform*. 2016;94:81-90. doi:10.1016/j.ijmedinf.2016.07.003

41. Sundström C, Gajecki M, Johansson M, Blankers M. Guided and Unguided Internet-Based Treatment for Problematic Alcohol Use – A Randomized Controlled Pilot Trial. 2016;8:1-19. doi:10.1371/journal.pone.0157817

42. Watson A, Bickmore T, Cange A, Kulshreshtha A. An Internet-Based Virtual Coach to Promote Physical Activity Adherence in Overweight Adults : Randomized Controlled Trial Corresponding Author : 14. doi:10.2196/jmir.1629

43. Santarossa S, Kane D, Senn CY, Woodruff SJ. Exploring the Role of In-Person Components for Online Health Behavior Change Interventions: Can a Digital

Person-to-Person Component Suffice? *J Med Internet Res*. 2018;20(4):e144. doi:10.2196/jmir.8480

44.   Schueller SM, Tomasino KN, Mohr DC. Integrating Human Support Into Behavioral Intervention Technologies: The Efficiency Model of Support. *Clin Psychol Sci Pract*. 2017;24(1):27-45. doi:10.1111/cpsp.12173

45.   Andersson G, Cuijpers P. Internet-Based and Other Computerized Psychological Treatments for Adult Depression: A Meta-Analysis. *Cogn Behav Ther*. 2009;38(4):196-205. doi:10.1080/16506070903318960

46.   Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: A systematic review of features and content quality. *J Med Internet Res*. 2015;17(8):e198. doi:10.2196/jmir.4581

47.   Nikolaou CK, Lean MEJ. Mobile applications for obesity and weight management: current market characteristics. *Int J Obes (Lond)*. Published online November 22, 2016. doi:10.1038/ijo.2016.186

48.   Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. *J Med Internet Res*. 2011;13(1):e30. doi:10.2196/jmir.1602

49.   David H. Barlow, Shannon Sauer-Zavala, Kristen K. Ellard, et al. *Unified Protocol for Transdiagnostic Treatment of Emotional Disorders*. Oxford University Press; 2017. Accessed July 8, 2020. https://www.google.com/books/edition/Unified_Protocol_for_Transdiagnostic_Tre /0cM9DwAAQBAJ?hl=en&gbpv=1&dq=unified+protocol+clair+robbins&printse c=frontcover

50.   Barlow DH, Ellard KK, Fairholme CP, et al. *Unified Protocol for Transdiagnostic Treatment of Emotion Disorders Workbook*. (Barlow DH, ed.). Oxford University Press; 2011.

51.   National Center for Education Statistics. *Literacy, Numeracy, and Problem Solving in Technology-Rich Environments Among U.S. Adults: Results from the Program for the International Assessment of Adult Competencies 2012*.; 2012.

52.   Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. *https://doi.org/101080/10447310802205776*. 2008;24(6):574-594. doi:10.1080/10447310802205776

53.   Short CE, DeSmet A, Woods C, et al. Measuring engagement in eHealth and mHealth behavior change interventions: Viewpoint of methodologies. *J Med Internet Res*. 2018;20(11):e292. doi:10.2196/jmir.9397

54. Dennis JP, Vander Wal JS. The cognitive flexibility inventory: Instrument development and estimates of reliability and validity. *Cognit Ther Res*. 2010;34(3):241-253. doi:10.1007/s10608-009-9276-4

55. Feldman G, Westine M, Edelman A, Higgs M, Renna M, Greeson J. Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). *Handb Assess Mindfulness Res*. Published online 2022:1-24. doi:10.1007/978-3-030-77644-2_19-1

56. Litman JA. The COPE inventory: Dimensionality and relationships with approach- and avoidance-motives and positive and negative traits. *Pers Individ Dif*. 2006;41(2):273-284. doi:10.1016/j.paid.2005.11.032

57. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the Overall Depression Severity and Impairment Scale. *Psychol Assess*. 2014;26(3):815-830. doi:10.1037/a0036216

58. Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an Overall Anxiety Severity And Impairment Scale (OASIS). *Depress Anxiety*. 2006;23(4):245-249. doi:10.1002/DA.20182

59. Bowen DJ, Kreuter M, Spring B, et al. How We Design Feasibility Studies. *Am J Prev Med*. 2009;36(5):452-457. doi:10.1016/J.AMEPRE.2009.02.002

60. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. 2004;10(2):307-312. doi:10.1111/J..2002.384.DOC.X

61. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res*. 2011;45(5):626-629. doi:10.1016/j.jpsychires.2010.10.008

62. Pearson N, Naylor PJ, Ashe MC, Fernandez M, Yoong SL, Wolfenden L. Guidance for conducting feasibility and pilot studies for implementation trials. *Pilot Feasibility Stud*. 2020;6(1):1-12. doi:10.1186/S40814-020-00634-W/FIGURES/2

63. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: The what, why and how. *BMC Med Res Methodol*. 2010;10(1):1-10. doi:10.1186/1471-2288-10-1/TABLES/3

64. Kannan S, Gowri S. Pilot studies: Are they appropriately reported? *Perspect Clin Res*. 2015;6(4):207. doi:10.4103/2229-3485.167097

65. Kirsch I. Response expectancy theory and application: A decennial review. *Appl Prev Psychol*. 1997;6(2):69-79. doi:10.1016/S0962-1849(05)80012-5

66. Parker C, Scott S, Geddes A. Snowball Sampling. *SAGE Res Methods Found*. Published online 2021. Accessed June 21, 2022. http://dx.doi.org/9781529749946

67. Census US. *U.S. Census Bureau QuickFacts: North Carolina; United States*.; 2020. Accessed June 21, 2022. https://www.census.gov/quickfacts/fact/table/NC,US/PST045221

68. Gallup. *What Percentage of Americans Are LGBT?*; 2021. Accessed June 21, 2022. https://news.gallup.com/poll/332522/percentage-americans-lgbt.aspx

69. ResearchMatch. ResearchMatch Metrics. Published June 2022. Accessed June 21, 2022. https://www.researchmatch.org/metrics/

70. Wurm M, Klein Strandberg E, Lorenz C, et al. Internet delivered transdiagnostic treatment with telephone support for pain patients with emotional comorbidity: a replicated single case study. *Internet Interv*. 2017;10:54-64. doi:10.1016/J.INVENT.2017.10.004

71. Rondung E, Ternström E, Hildingsson I, et al. Comparing Internet-Based Cognitive Behavioral Therapy With Standard Care for Women With Fear of Birth: Randomized Controlled Trial. *JMIR Ment Heal*. 2018;5(3). doi:10.2196/10420

72. Strandberg CL, Klein E. An Internet-delivered version of the Unified Protocol as a secondary intervention for individuals with persistent pain and co-morbid emotional problems: A pilot study. *Unpubl Master's thesis, Örebro Univ Sweden*. 2015;1.

73. Reynolds GO, Saint-Hilaire M, Thomas CA, Barlow DH, Cronin-Golomb A. Cognitive-Behavioral Therapy for Anxiety in Parkinson's Disease. *Behav Modif*. 2020;44(4):552-579. doi:10.1177/0145445519838828

74. Tonkin-crine S, Bishop FL, Ellis M, Moss-morris R. Exploring Patients ' Views of a Cognitive Behavioral Therapy-Based Website for the Self-Management of Irritable Bowel Syndrome Symptoms Corresponding Author : 15:1-13. doi:10.2196/jmir.2672

75. Muñoz RF, Bunge EL, Chen K, et al. Massive Open Online Interventions : A Novel Model for Delivering Behavioral- Health Services Worldwide. Published online 2016. doi:10.1177/2167702615583840

Biography

Jack Brooks received a Bachelor of Arts in psychology from New York University in 2014 and a Master of Arts in clinical psychology from Duke University in 2019; this dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in clinical psychology from Duke University. Jack has been a coauthor on five published, peer-reviewed articles: Designing Ruby: Protocol for a 2-Arm, Brief, Digital Randomized Controlled Trial for Internalized Weight Bias (2021), Delivering bad news: A randomized trial of virtual vs. real standardized patients (2020), Consistent self-monitoring in a commercial app-based intervention for weight loss: results from a randomized trial (2019), Comparing self-monitoring strategies for weight loss in a smartphone app: A randomized controlled trial (2019), and Strategies used by older asthmatics for adherence to inhaled corticosteroids (2014), as well as a chapter ("Third-wave cognitive behavioral therapies for the treatment of anxiety among children and adolescents") in the book Pediatric Anxiety Disorders (2019). Jack is a member of the Society of Duke Fellows. She has received the four-year Dean's Graduate Fellowship at Duke University (2016-2020) and Race and the Professions Fellowship (2020-2021). Her research has been supported by a Charles Lafitte Foundation Graduate Grant Award for Psychological Research related to Identity, Diversity, Inclusion, Equity, and Thriving (2020), the Stella Powell-Williams Grant for Research Impacting Community (2020), and the Claire Hamilton Graduate Studies Travel Award (2016).