NORTHWESTERN UNIVERSITY

Mobile Apps for the Treatment of Depression

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

For the degree

DOCTOR OF PHILOSOPHY

Field of Clinical Psychology

By

Elizabeth Colleen Stiles-Shields

EVANSTON, ILLINOIS

September 2016

ABSTRACT

Mobile Apps for the Treatment of Depression

Colleen Stiles-Shields

The National Institute of Mental Health and the Agency for Healthcare Research and Quality convened an expert panel to identify the top research priorities in the field of behavioral intervention technology (BIT) research. The panel determined that if BIT research is to evolve in an effective way to reach and serve those with mental health needs, theoretical and research paradigms from multiple disciplines must be integrated and refined. The projects from the present research reflect this recommendation, linking usability testing methods and psychological theory to evaluate apps for depression. Apps are ideally suited as a new delivery mechanism to overcome barriers to depression interventions; however, there are many gaps in the literature regarding use of this delivery mechanism. The present research increased the knowledge of apps for depression by identifying: barriers to this delivery mechanism that implicate design changes, the learnability and learning performance of users following initial use, and the feasibility of evaluating use and outcomes of apps for depression. Usability testing and RCTs are forms of research typically conducted and consumed by different fields. However, in using both forms of research, the design, development, and deployment of BITs can be improved to reach those with depression.

Acknowledgements

This dissertation was literally completed through the sacrifice of blood, sweat, and teeth. However, it would not have been possible without the support of a wonderful community of people.

First, thank you to my committee. To David, not only for signing on for the years of work that is taking on a graduate student, but for choosing to be a mentor throughout that time. To Enid, for your persistence in getting me to ask well-informed questions, and to build from there. To Mary, for maintaining such a wonderful humor, but never laughing at my statistical questions.

Thank you to the Center for Behavioral Intervention Technologies. For building the apps, supporting all of my questions, and cheering on this entire project. To Chris, Ellen, Emily, Hannah, Hendricks, Jenna, Joyce, Kate, Kristina, Lauren, Mark, Marya, Nameyeh, Stephen, Susan, and the rest of CBITs at large. Thank you to the National Institute of Mental Health, for funding my time during the execution of this dissertation.

Thank you to my family and friends. For tolerating my stress, distracting me, and being magnificent. To my parents, for always supporting my education and giving me the mentality to shake things off and keep running against the wind. To my brothers, for being the best examples and most loving antagonizers. To the Stiles, Shields, and Begales; especially, Dominic and Audrey, for being my study buddies and favorite distractions. To Becca, Em, Jamie, Lauren, and Sara, for being the most fearsome, loyal, and wonderful women I have met.

Finally, thank you to Joe. You selflessly took on my dreams as your own, and your unshakable faith in where we are headed made this all seem possible.

List of Abbreviations

ANOVA = Analysis of Variance

BA = Behavioral Activation

BDI = Beck Depression Inventory

BITs = Behavioral Intervention Technologies

CBITs = Center for Behavioral Intervention Technologies

CBT = Cognitive Behavioral Therapy

CT = Cognitive Therapy

CTAS = Cognitive Therapy Awareness Scale

DSM-IV = Diagnostic and Statistical Manual-IV

FAQs = Frequently Asked Questions

ICD-10 = International Statistical Classification of Diseases and Related Health Problems

IRB = Institutional Review Board

MAD = Mobile Apps for Depression

MDD = Major Depressive Disorder

MINI = Mini International Neuropsychiatric Interview

PHQ-9 = Patient Health Questionnaire-9

QIDS = Quick Inventory of Depressive Symptoms

RCTs = Randomized Controlled Trials

REDCap = Research Electronic Data Capture

SUS = System Usability Scale

Dedication

To Elizabeth Connelly Stiles, the trail you continue to blaze has started a fire reaching farther than you will ever know.

Table of Contents

List of Figures and Tables

Figures

Tables

Chapter I: Introduction

Background

Significance of Depression

Depressive disorders are the leading cause of disability in the United States (Murray & Lopez, 1996) and worldwide (Ferrari, Charlson, et al., 2013). Major depressive disorder (MDD) is common, occurring in 7.6% of Americans (Pratt & Brody, 2014), and imposes a very high societal burden in terms of cost, morbidity, suffering, and mortality (Ferrari, Charlson, et al., 2013; Wells et al., 2002; Whooley & Simon, 2000; Whooley, Stone, & Soghikian, 2000). However, even depression that does not meet the full diagnostic criteria (subthreshold depression) causes distress, disability, and perceived health problems comparable to that of MDD (Backenstrass et al., 2005; Cuijpers, Vogelzangs, et al., 2013; Judd, Paulus, Wells, & Rapaport, 1996; Rucci et al., 2003). Subthreshold depression is more common than MDD, with prevalence rates estimated between 13.8% and 20.9% (Kessler, Zhao, Blazer, & Swartz, 1997; Laborde-Lahoz et al., 2015; Rucci et al., 2003; Zung, Broadhead, & Roth, 1993). The prevalence and debilitating nature of depressive disorders make them a prime target for intervention in the general population.

Traditionally Delivered Interventions for Depression

Effective Psychological Treatments. Psychological interventions for depression are effective (Cuijpers, van Straten, Andersson, & van Oppen, 2008) and desirable to patients (Bedi et al., 2000; Brody, Khaliq, & Thompson, 1997; Churchill et al., 2000; Dwight-Johnson, Sherbourne, Liao, & Wells, 2000; Priest, Vize, Roberts, Roberts, & Tylee, 1996). Behavioral and cognitive interventions carry the strongest body of evidence for the treatment of depression (Butler, Chapman, Forman, & Beck, 2006; Cuijpers, Berking, et al., 2013; Cuijpers, van Straten,

& Warmerdam, 2007; Dobson, 1989). Behavioral interventions (e.g., Behavioral Activation) target activity and mood monitoring, with an emphasis on increasing activities that instill a sense of pleasure and/or accomplishment. This sense of reward from rewarding behaviors positively impacts mood and reinforces future engagement in these behaviors (Martell, Dimidjian, & Herman-Dunn, 2010). Cognitive interventions (e.g., Cognitive Therapy) focus upon thought restructuring, a process involving the identification and appraisal of maladaptive thoughts and creating adaptive counter thoughts. The shift in focus from thoughts that tend to negatively impact mood, and focusing on more positive and realistic thoughts positively impacts mood over time (J. S. Beck, 1995). Behavioral and cognitive interventions for depression have been traditionally delivered via face-to-face administration (J. S. Beck, 2011).

Effective Psychological Treatments and Learning. Behavioral and cognitive interventions focus on educating patients regarding the impact of their thoughts and/or behavior on their mood (J. S. Beck, 2011). Engaging patients in monitoring and exploration via Socratic questioning is intended to build insight into opportunities for change and ultimate symptom reduction. Learning and application of intervention skills are therefore noted to be among the possible mechanisms supporting symptom change in behavioral and cognitive interventions (Barber & DeRubeis, 1989; Hundt, Mignogna, Underhill, & Cully, 2013). Relatedly, gains in learning intervention concepts and skills have been found to significantly predict depression outcomes in these interventions (Jarrett, Vittengl, Clark, & Thase, 2011; Miner, Schueller, Lattie, & Mohr, 2015). Given the association of learning with symptom reduction, explorations of means to increase learning in behavioral and cognitive interventions have increased in recent years (Gumport, Williams, & Harvey, 2015; Harvey et al., 2014). While multiple factors have been implicated in symptom change in depression treatments, learning is frequently considered a

key mechanism. Methods of enhancing learning in interventions for depression are therefore being explored to support better outcomes.

Barriers to Traditional Treatment Delivery. While efficacious treatments exist, most adults with depression do not receive psychological interventions (Gonzalez et al., 2010). A number of barriers to the delivery of face-to-face psychological interventions have been identified. First, there is an inadequate work force of mental health providers trained in evidence-based treatments to meet the needs of the number of people with depression (Addis & Krasnow, 2000; Karekla, Lundgren, & Forsyth, 2004; Thomas, Ellis, Konrad, Holzer, & Morrissey, 2009; Weil, 2015). Indeed, with 21-30 million Americans requiring treatment for depression annually, the current work force is unable to meet such a demand with standard one-on-one intensive treatments (Kazdin & Blase, 2011). Further, the available work force is also dispersed inequitably, such that there tend to be more professionals in urban areas than rural (Thomas et al., 2009). Second, symptoms associated with depression, such as low motivation and hopelessness, impact initiation and adherence to treatment (American Psychiatric Association, 2013; DiMatteo, Lepper, & Croghan, 2000; Mohr et al., 2010). Depression is also highly comorbid with anxiety disorders (Fava et al., 2000; Sartorius, Üstün, Lecrubier, & Wittchen, 1996), which can exacerbate these symptom-based barriers (e.g., having anxiety around "what others might think" if seeking treatment; Olfson et al., 2000). Third, cultural and stigma barriers have been identified as impacting beliefs about treatment, help-seeking behavior, and engagement with treatment (Cooper et al., 2003; Gonzalez et al., 2010; Menke & Flynn, 2009; Möller-Leimkühler, 2002). Fourth, practical barriers, such as time, travel, childcare, etc., are known barriers to depression treatment (Mohr et al., 2006; Mohr et al., 2010; Wallace, Weeks, Wang, Lee, & Kazis, 2006). Despite the existence of effective interventions for depression,

identified barriers to initiation and maintenance of face-to-face interventions highlight the need for alternative delivery approaches.

Behavioral Intervention Technologies (BITs) for Depression

To address these barriers and extend care capacity, Behavioral Intervention Technologies (BITs) are being integrated into multiple healthcare systems (Christensen & Hickie, 2010; Darkins et al., 2008; National Institute for Clinical Excellence, 2004). BITs are the use of technologies to assist in making and sustaining behavior changes related to health, mental health, and general wellness (Mohr, Burns, Schueller, Clarke, & Klinkman, 2013). Common examples of BITs include smartphone health apps, treatment and prevention websites, sensors used in activity trackers, and smartwatches. The term BIT is used for its specificity towards technology-supported behavior change. Indeed, the terms eHealth and mHealth encompass greater aspects of medicine and informatics than behavior change (e.g., electronic health records; Oh, Rizo, Enkin, & Jadad, 2005), whereas the term BIT focuses solely on technologies that support behavior change (Burns & Mohr, 2013). There is a small, but growing, body of evidence supporting the use of evidence-based interventions delivered via technology (Clarke & Yarborough, 2013). Through the use of multiple technology platforms, BITs offer the possibility of extending access to treatment and prevention interventions for a variety of health and wellness concerns, including depression (Mohr, Burns, et al., 2013).

Efficacy and Human Support in Web-Based BITs. The majority of evaluated BITs to date have been delivered via web-based platforms (i.e., web-based BITs; Barak, Klein, & Proudfoot, 2009; Burns & Mohr, 2013; Mohr, Burns, et al., 2013). Meta-analyses suggest efficacy of web-based BITs in ameliorating depressive symptoms (Andersson & Cuijpers, 2009; Andrews, Cuijpers, Craske, McEvoy, & Titov, 2010; Spek et al., 2007). Further, those

interventions including human support demonstrate increased efficacy and retention compared to unsupported interventions (Andersson & Cuijpers, 2009; Richards & Richardson, 2012). However, while questioned for the quality of support provided (Jones et al., 2015), not all human supported interventions have consistently demonstrated improvement, compared to typical care for depression (Gilbody et al., 2015). Additionally, the benefits of human support in web-based depression BITs may be dependent on symptom severity, such that those with mild depression experience gains from human support (Newman, Szkodny, Llera, & Przeworski, 2011). Despite some mixed evidence, web-based BITs appear to be efficacious in the treatment of depression and these effects may be benefited by the use of human support.

Adherence and Barriers to Web-Based BITs. Adherence to web-based BITs evaluated via randomized controlled trials (RCTs) has ranged from 50-70% (Christensen, Griffiths, & Farrer, 2009). Predictors of adherence have varied (Beatty & Binnion, 2016), but include less severe baseline depression, being younger, being female, higher beliefs in treatment success, having guidance, and having lowered knowledge of psychological interventions for depression (Beatty & Binnion, 2016; Christensen et al., 2009). However, open access web-based BITs demonstrate much lower adherence than RCT participation (Christensen et al., 2009; Christensen, Griffiths, & Korten, 2002; Christensen, Griffiths, Korten, Brittliffe, & Groves, 2004; Donkin et al., 2011; Melville, Casey, & Kavanagh, 2010). Not surprisingly, different barriers associated with web-based BITs have been identified. These include needing to have access to and be in front of a computer with broadband internet access, lack of time, technical malfunctions, and finding content impersonal or irrelevant (Beatty & Binnion, 2016; Rainie & Cohn, 2014). Given these barriers, other delivery platforms for BITs are being developed and evaluated.

Mobile Apps for Depression

The marketplace for smartphone applications (apps) has increased exponentially in recent years, with over 165,000 health and wellness apps publically available in 2015. The majority of these apps focus on wellness (i.e., fitness, stress, diet), with subsets targeting disease-specific topics. Mental health concerns, such as depression, are the most common focus of disease-specific apps (IMS Institute for Healthcare Informatics, 2015). Most apps targeted towards users with depression have a single functionality, utilizing the dimensions of: informing, instructing, recording, displaying, guiding, alerting, or communicating with users (IMS Institute for Healthcare Informatics, 2015). For the purposes of the present discussion, apps for depression are broadly defined as the delivery of psychoeducation and/or an intervention skill via a smartphone app.

Apps are ideally suited as a new delivery mechanism to overcome face-to-face and web-based barriers to depression interventions. First, given their instantiation within a mobile device that users tend to keep with them throughout the day, apps provide opportunities for real-time monitoring, assessment, and interventions in the real-world conditions of an individual with depression (Proudfoot, 2013). This increases accessibility, but also the likelihood of accurate assessment of symptoms or impact of an in-the-moment intervention (i.e., patients with depression may otherwise provide inaccurate reports when assessed at later times, due to known impairments associated with depression, such as decreased attention and memory (Behnken et al., 2010; Campbell & Macqueen, 2004; Lee, Hermens, Porter, & Redoblado-Hodge, 2012; MacQueen et al., 2003; Videbech & Ravnkilde, 2004)). Second, their instantiation within smartphones increases reach. Nearly two-thirds of all Americans own smartphones (Smith, 2015). Furthermore, there is a growing number of smartphone-dependent users, which is defined

as: 1) owning a smartphone, 2) not having broadband internet access at home, and 3) having limited abilities to access the internet outside of a smartphone (Smith, 2015). Smartphone-dependency is particularly prevalent among younger adults, minorities, and low income users (Pew Research Center, 2014; Smith, 2015), groups which are likely to experience stigma-related or practical barriers to traditional treatment delivery (Cooper et al., 2003; Gonzalez et al., 2010; Menke & Flynn, 2009; Mohr et al., 2006; Mohr et al., 2010; Möller-Leimkühler, 2002; Wallace et al., 2006). Finally, there appears to be interest in specialty apps focusing on specific health and mental health needs. In a recent national survey, nearly 60% of those with mobile phones have downloaded and used a health-related app (Krebs & Duncan, 2015). Apps targeting mental health have been found to be of interest to populations with mental health concerns (Torous, Friedman, & Keshavan, 2014). Given the proliferation of smartphones and the increase in accessibility and opportunities for real-time interactions, apps are an incredibly promising delivery mechanism to overcome barriers to depression interventions.

Problems and Gaps in Knowledge. While many apps for depression are available, clinical research, by its nature, lags behind the advancement of this technology (Mohr, Cheung, Schueller, Hendricks Brown, & Duan, 2013). Indeed, despite a growing body of evidence for their promise and efficacy (Donker et al., 2013; Shen et al., 2015), a number of gaps in the literature regarding apps for depression remain. Additionally, specific problems have been identified, which highlight needs for improvement in this delivery mechanism.

Poor Usage and Efficacy of Available Apps. Despite the growing number of apps geared toward the monitoring and treatment of depression, a recent review indicates that it is unlikely users are able to locate a reliable, credible, and evidence-based informed app(s) for depression (Shen et al., 2015). Relatedly, app use is low. Indeed, even when prescribed by a healthcare

provider, roughly 70% of users download a mental health app, with only 40% of these users sustaining use up to 30 days following download (IMS Institute for Healthcare Informatics, 2015). This is similar to other intervention apps evaluated in research trials. Recent reviews of mental health, and more broadly, health apps note that recent findings from trials involving intervention apps should be interpreted with caution given the small number of studies and participants, high risk of bias, and unknown efficacy of long-term follow-up (Donker et al., 2013; Payne, Lister, West, & Bernhardt, 2015). While initial evidence suggests behavioral and cognitive interventions delivered via apps may be as effective for users with depression as when delivered via computer (Watts et al., 2013), currently available apps for depression generally lack evidence regarding their efficacy. Further, even when based on psychological theory, there is no evidence that the instantiation of these theories developed for face-to-face treatment translates into apps (Buijink, Visser, & Marshall, 2013; Huckvale, Car, Morrison, & Car, 2012; Sucala et al., 2013). More research is required to establish the efficacy of depression interventions delivered via apps and to build the accessibility of evidence-based apps.

Unidentified Barriers. One possible contributing factor to the poor use and efficacy of apps for depression is barriers. While apps are being designed and disseminated due to their assumed accessibility, as a new delivery mechanism, users likely perceive and/or experience unique barriers to their use for depression. However, barriers to utilizing apps for depression have not been explored or identified. Ascertaining these barriers is critical to the success of future apps; as some barriers might be easily addressed through design, such as providing specific psychoeducation at download. However, without identification of such barriers, app researchers and designers must primarily rely on intuition (Riley et al., 2011). This methodology promotes a risk that design choices will create a mismatch with user needs for this delivery

mechanism, promoting continued poor uptake and use. Therefore, identification of user perceived barriers to apps for depression are needed to improve the design of future apps, with the aim of increasing use and uptake.

Unidentified Learning Processes and Successes. Another possible contributing factor to poor use and efficacy of depression apps is a poor understanding of learning processes and outcomes in apps. The majority of mental health apps are designed with the purpose of providing information regarding symptoms or their management (IMS Institute for Healthcare Informatics, 2015; Shen et al., 2015). As previously described, behavioral and cognitive interventions for depression utilize patients' learning about their internal (i.e., thoughts and mood) and external (e.g., behaviors) experiences as a key mechanism to enacting symptom change (Barber & DeRubeis, 1989). As apps aim to translate behavioral and cognitive strategies to users via technology, it is therefore not surprising that the majority of apps targeting mental health are designed to provide learning opportunities, ranging from providing didactic content to enabling practice of an intervention skill via use of a tool (Donker et al., 2013). However, little is known about how (i.e., processes) and what (i.e., symptom knowledge, intervention skill implementation, etc.) users learn from apps targeting mental health concerns. Evaluation of user learning processes and outcomes is needed to better understand the impact of learning in apps for depression.

Behavioral vs. Cognitive Approaches. Behavioral Activation (BA) and Cognitive Therapy (CT) interventions are generally believed to be equivalent when delivered face-to-face (Aderka, Nickerson, Boe, & Hofmann, 2012; Busch, Kanter, Landes, & Kohlenberg, 2006; Cuijpers et al., 2007; Dimidjian et al., 2006; Dobson et al., 2008; Hardy et al., 2005; Hunnicutt-Ferguson, Hoxha, & Gollan, 2012; Tang & DeRubeis, 1999; Tang, DeRubeis, Beberman, &

Pham, 2005; Vittengl, Clark, & Jarrett, 2005). The noted exception being that BA may outperform CT in severely depressed patients (Dimidjian et al., 2006). The majority of currently available apps are informed by principles and intervention strategies from BA and CT (Donker et al., 2013; Shen et al., 2015). However, the efficacy of BA and CT in face-to-face treatments cannot be assumed to remain consistent when delivered through a new medium (i.e., apps). It is unclear how BA and CT-informed apps for depression impact the symptoms of users. Additionally, to date, BA and CT have not been directly compared when delivered via apps for depression, and it is unclear if one might be better suited for use in this delivery mechanism. There is therefore a need for the evaluation of BA and CT-informed apps for depression, both for their individual and comparative efficacies in users with depression.

Problem Statement

Depression is a significant public health concern. Effective treatments for depression exist, and likely require learning as a means to target symptoms. While these treatments are available, multiple access and practical barriers prevent initiation or maintenance of traditionally delivered interventions. BITs, specifically those delivered via smartphone apps, are being increasingly developed and deployed as a means to overcome access barriers to depression treatment. However, the following gaps in knowledge exist in apps for depression. First, it is unclear what barriers users anticipate in using an app for depression. Second, apps are delivering interventions that require learning to enhance efficacy, yet it is unclear what processes (i.e., how) and outcomes (i.e., what) users learn from apps informed by interventions emphasizing evidence-based skills. Finally, usage and efficacy of behavioral and cognitive treatment apps are varied and poorly defined, individually and in comparison with one another.

Goals

The present research has three primary goals to address these gaps in knowledge: 1) Identify user perceived barriers to the use of apps for depression to inform future design decisions; 2) Evaluate the learnability and learning performance of users following initial use of an app for depression; and 3) Evaluate usage and impact on depressive symptoms following access to a BA or CT-informed app.

Research Question 1: What are the primary barriers to the initiation and maintenance of use of an app for depression?

Research Question 2: How might learning be defined and evaluated in an app for depression?

Research Question 3: What is the overall usage and effect on depressive symptoms for a BA and a CT-informed mobile app? Does a BA or CT-informed app impact symptoms of depression compared to a waitlist control group or each other?

Approach

The problems, goals, and approaches of the present research are outlined in Table 1. The research goals were approached via two research methodologies, usability testing and RCTs, detailed below.

Usability Testing

Usability testing is a method of evaluation that involves testing users' interactions with a product and system to improve design and to ensure that a technology is intuitive and easy to use. Sometimes confused with informal inquiry of user opinions of a product, usability testing requires systematic observation of a planned task or scenario carried out by an actual or potential user (Usability.gov). In-lab usability testing (i.e., a session requiring participants to attend in

person) included the participation of 20 adults. As depression is a condition that is frequently chronic, characterized by patterns of remissions and relapses (Judd, Paulus, & Zeller, 1999; Mueller et al., 1999; Paykel, 2008), equal numbers of participants currently above and below the criteria for a referral for psychotherapy were recruited (The MacArthur Foundation Initiative on Depression and Primary Care, 2004). This sampling ensured that the goals of this study were being measured with likely end users, ranging from those with no or mild depressive symptoms (subthreshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score less than 10) to those with moderate or severe depressive symptoms (threshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score greater than or equal to 10; Kroenke & Spitzer, 2002).

Identify User Perceived Barriers to the Use of Apps for Depression. To achieve the first goal (detailed in Chapter II), a card sorting task that ranked and grouped barriers to use of apps for depression was completed. Card sorting tasks are designed as a means to categorize and organize variables and ideas (Nielsen, 2004) and are commonly utilized to inform multiple design processes and decisions (Wood & Wood, 2008). Participants first completed a card sorting task identifying barriers to face-to-face treatment, as a primer to identification of barriers. Participants then completed a card sorting task identifying mobile barriers.

Aims. The card sorting tasks were exploratory in nature, so no hypotheses were made. The aims of completing the card sorting tasks therefore were to: 1) identify perceived barriers to depression treatment through a mobile app and 2) identify overlap in primary barriers for mobile app treatment with traditional treatment delivery barriers.

Evaluate the Learnability and Learning Performance of Users Following Initial Use of an App for Depression. To achieve the second goal (detailed in Chapter III), usability testing

methods were used to measure the usability attributes of learnability (i.e., the level of ease through which a user gains proficiency with a technology) and learning performance (i.e., actual impact of interaction with a technology on performance of a task/acquisition of knowledge; Nielsen, 1993; Wood & Wood, 2008). The app for depression evaluated was Thought Challenger, an app informed by CT (Lattie et al., In Press). The design and execution of the testing of Thought Challenger followed an established framework for evaluating apps (Zhang & Adipat, 2005).

Aims. The testing was exploratory in nature and therefore no hypotheses were made. The aims were to address three usability questions to evaluate the efficacy of the app (i.e., Thought Challenger) in achieving the intended learning objectives: 1) How well does a user initially interact with the Thought Challenger app without instruction; 2) Is a user able to learn the skill of cognitive restructuring from the app; and 3) Does use of Thought Challenger change baseline knowledge of cognitive therapy elements.

Randomized Controlled Trial

The Mobile Apps for Depression (MAD) Trial included the participation of 30 adults. Participants were randomized to one of three conditions: 1) Boost Me, 2) Thought Challenger, or 3) Waitlist Control. Boost Me is a native app (i.e., an app downloaded directly to a mobile device and designed to work with a specific operating system) which instantiates activity scheduling, a core strategy of BA, which aims to increase rewarding activities and monitoring of mood in relation to behavior (Cuijpers et al., 2007; Martell et al., 2010). Thought Challenger is a native app which instantiates thought restructuring, the core strategy in CT that involves identifying and appraising maladaptive thoughts and creating adaptive counter thoughts (J. S. Beck, 2011; Lattie et al., In Press). Boost Me and Thought Challenger participants received six weeks of weekly

coaching sessions, didactic content delivery (one lesson per week), and use of the app. The waitlist control group did not receive any intervention until the passage of 10 weeks occurred, to account for both the intervention period (six weeks) and follow-up period (four weeks).

Evaluate Usage and Impact on Depressive Symptoms Following Use of a BA or CT-Informed App. To achieve the third goal (detailed in Chapter IV), an RCT was conducted, with participants randomized to either Boost Me, Thought Challenger, or Waitlist Control on a 1:1:1 ratio, with a block size of 4. The primary outcome, depressive symptomology, was measured using the PHQ-9 at baseline, week 3 (mid-treatment), week 6 (end of treatment), and week 10 (one month follow-up). Usage was defined by number of times the app was launched, events or thoughts were logged, and the review function was launched during the treatment period (i.e., six weeks).

Aims. As the sample size of the trial was not powered for hypothesis testing, the aims of this trial were as follows. First, conduct pilot feasibility of an RCT of utilizing single apps focusing on a discrete behavioral skill for depression. Second, identify usage and symptom response trends between the Boost Me and Thought Challenger apps, and symptom trends compared with a waitlist control group.

Different Types of Research: Usability Testing and Randomized Controlled Trials

Usability testing and RCTs are two forms of research that each possesses a variety of strengths and limitations. To conclude the present work, Chapter V outlines the two research methods. Given the increasing use of both in the development and evaluation of BITs, benefits and challenges of working across disciplines are also described.

Conclusions

While BITs have the capacity to reach a wide variety of users, actual use tends to be poor for the majority of BITs (Clarke & Yarborough, 2013; IMS Institute for Healthcare Informatics, 2015; Maher et al., 2014). Initial evidence suggests that active users of publically deployed BITs are likely those who have already enacted behavioral change and are utilizing BITs as monitoring and maintenance tools (IMS Institute for Healthcare Informatics, 2015). An illustration of this phenomenon is the number of active users of the self-monitoring diet tracker, "The Eatery," who categorized themselves at the onset as already maintaining a "strict diet" (Helander, Kaipainen, Korhonen, & Wansink, 2014). Indeed, such users who already have enacted behavioral change tend to fall in the health and wellness domain of BITs (IMS Institute for Healthcare Informatics, 2015), highlighting the need to specifically target potential BIT users with mental health needs, such as depression. Recognizing this need, the National Institute of Mental Health and the Agency for Healthcare Research and Quality convened an expert panel to identify the top research priorities in the field of mental health BIT research. The panel determined that if BIT research is to evolve in an effective way to reach and serve those with mental health needs, theoretical and research paradigms from multiple disciplines must be integrated and refined (Mohr, Burns, et al., 2013).

Stemming from this recommendation, the proposed projects link psychological theory and usability testing to guide the evaluation of apps for depression. This will be the first: 1) use of a card sorting task to identify user perceived barriers to the use of apps for depression; 2) use of an established framework to design and implement usability testing methods of a mobile app designed to increase user learning of a discrete behavioral strategy in users with depression; and 3) comparison of two basic psychological principles, BA and CT, instantiated in a mobile app,

compared to waitlist control. The results of the proposed projects will provide the first evaluation of mobile barriers, learning of an intervention strategy in an app, and insight into the efficacy and use of mobile app methods targeting increasing positive activities and thought restructuring for depressive symptoms.

Chapter II: Identifying Barriers to the Use of Apps for Depression

Introduction

Depressive disorders are the leading cause of disability worldwide (Ferrari, Charlson, et al., 2013). While efficacious treatments for depression exist (Cuijpers et al., 2008), practical and emotional access barriers interfere with initiation and maintenance of face-to-face (i.e., traditionally delivered) treatments (Mohr et al., 2006). Therefore, to address this mental health epidemic, significant changes must be made in the strategy with which interventions are delivered. To extend care capacity, technologies are being integrated into multiple health care systems as a delivery mechanism for behavioral health interventions (Christensen & Hickie, 2010; Darkins et al., 2008; National Institute for Clinical Excellence, 2004). The use of web-based delivery platforms have demonstrated efficacy across a broad range of mental health outcomes (Andrews et al., 2010; Mohr, Burns, et al., 2013), however, barriers to this delivery method, such as needing to be in front of a computer, impact uptake and usage (Renton et al., 2014). Consequently, a growing body of research is examining the use of smartphones, which offer the potential to provide a nearly continuous connection between a care system and patients, to deliver interventions.

As smartphones grow in popularity, their ability to serve as a delivery mechanism for behavioral health interventions with the potential to reach increasingly broad communities increases. Indeed, a growing number of people are becoming smartphone-dependent (Pew Research Center, 2014; Smith, 2015). Smartphone-dependency is defined as owning a smartphone, not having broadband internet access at home, and having limited abilities to access the internet outside of a smartphone (Smith, 2015). Through their instantiation in smartphones, apps are ideally suited to be accessed by users in real-time and in real-world conditions

(Proudfoot, 2013), likely overcoming many previously identified barriers to interventions delivered via face-to-face and computers (Mohr et al., 2006; Renton et al., 2014). However, while apps may address many barriers to other delivery mechanisms, they likely have unique barriers of their own. Identifying these barriers is critical to the success of future iterations of apps in delivering care to those with depression, particularly for those likely to face substantial known barriers to accessing traditionally delivered care.

Identification of barriers might implicate changes in the design of apps. For example, if concerns regarding efficacy of an app in addressing psychological symptoms is a barrier, design could shift to include providing specific psychoeducation at download related to efficacy. However, without identification of such barriers, app designers must primarily rely on intuition (Riley et al., 2011). This promotes a risk that design choices will create a mismatch with user needs or perceptions for this delivery mechanism. Identification of barriers may therefore improve the information available for those design and development of apps.

The means to identify barriers to the use of apps for depression may include a number of strategies, ranging from self-report questionnaires to moderated focus groups. However, a methodology that has been commonly used to inform multiple design processes and decisions is a card sorting task (Wood & Wood, 2008). Card sorting tasks are designed as a means to categorize and organize variables and ideas (Nielsen, 2004). Card sorting therefore enables the identification of potential end users' perception of barriers to the use and uptake of apps for depression. To our knowledge, card sorting tasks have not been used as a means to identify barriers to apps.

The purpose of the current study is to identify user perceived barriers to initiation and maintenance of apps for depression. The aims of completing the card sorting tasks therefore are

to: 1) identify perceived barriers to depression interventions delivered via apps and 2) identify overlap in primary barriers for intervention delivery via apps with traditional delivery methods (i.e., face-to-face) barriers.

Methods

Procedure

Recruitment of participants occurred from July to August 2015 from online postings in Chicago and nearby areas, resulting in the participation of 20 adults. Current recommendations for a card sorting task sample size is 15 (Nielsen, 2004), making the sample of 20 sufficient for the present study. Inclusion criteria were: being at least 18 years of age, the ability to attend an in-lab session, and ability to speak and read in English. As depression is a condition that is frequently chronic and characterized by patterns of remissions and relapses (Judd et al., 1999; Mueller et al., 1999; Paykel, 2008), equal numbers of participants currently above and below the criteria for a referral for psychotherapy were recruited (The MacArthur Foundation Initiative on Depression and Primary Care, 2004). This sampling ensured that perceived barriers were being measured with likely end users, ranging from those with no or mild depressive symptoms (subthreshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score less than 10) to those with moderate or severe depressive symptoms (threshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score greater than or equal to 10; Kroenke & Spitzer, 2002). Participants who completed the card sorting task, as well as an in-lab usability testing session (see Chapter III), were compensated $20 in petty cash for their time and participation. In compliance with the University's Institutional Review Board (IRB), participants completed an online screening consent prior to the collection of any data and were consented in-person for the card sorting and usability testing session.

Card Sorting. To identify barriers to use and engagement with apps that are specific to users with depression, two sort card sorting tasks using open sort methods were employed. Open card sorting refers to providing participants topics and asking them to sort them into groups that make sense to them, as opposed to a closed card sorting in which the topics would be organized into predefined groups (Wood & Wood, 2008). The first card sort was related to barriers to face-to-face delivery of interventions for depression and the second was related to barriers to app delivery of interventions for depression. This order was chosen, as a concern was that if participants asked to consider barriers to an app, they might not be familiar with the concept of an intervention app. If so, participants might identify barriers solely related to phone functionality (e.g., battery) or commonly used apps (e.g. Facebook). However, people are generally able to identify barriers to face-to-face interventions and having participants first consider these barriers promotes thinking regarding intervention barriers as well. Barriers listed for both tasks were informed by findings from the literature and polls from content experts at the Center for Behavioral Intervention Technologies (CBITs; Mohr et al., 2006; Mohr et al., 2010).

Prior to each card sorting task, participants were read the following prompt:

> I'm providing you with a stack of cards that have reasons that people might not want to or be able to (card sort 1: attend face-to-face therapy/card sort 2: use a mobile app for treatment) when feeling down. I would like you to go through the cards and choose the ones you think are barriers to (card sort 1: attending face-to-face therapy/card sort 2: using a mobile app for treatment). Once you choose them, please decide which ones are the biggest barriers. As you can see, the table is labeled to help you put ideas down from biggest barriers to smallest. You might notice that some overlap into groups in your mind; feel free to put them into groups. If there are cards you think do not apply, feel free

to put them over here to be discarded. If there are cards with reasons missing, we can add more (indicate blank cards and marker). Please feel free to think aloud as you go through the cards.

The card sorting tasks were timed and audio recorded, and photographs of the completed tasks were taken to ensure the moderator recorded the groupings correctly. Participants were provided time to supply a rationale for their choices following the tasks. This qualitative data was intended to enrich the findings and aid in the interpretation of groupings. The stacks of cards were shuffled between participants to remove any possible bias from rankings of other participants.

Measures

Study data were collected and managed using Research Electronic Data Capture (REDCap) electronic data capture tools hosted at Northwestern University (Harris et al., 2009). REDCap is a secure, web-based application designed to support data capture for research studies, providing 1) an intuitive interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for importing data from external sources.

At screening, participants were asked to provide demographic information (i.e., gender, race/ethnicity, age, education, and employment status). Further, they completed the Patient Health Questionnaire-9 (PHQ-9), a 9-item self-report instrument measuring depressive symptomology with scores ranging from 0-27 (Kroenke & Spitzer, 2002). Participants below the criteria for a referral to psychotherapy were defined as having a PHQ-9 score as below 10 (i.e., with no to mild depressive symptoms), whereas those meeting criteria for a referral to psychotherapy were defined as having a score of 10 or greater (i.e., with moderate to severe

depressive symptoms). This criteria reflects the MacArthur recommendations for referrals to psychotherapy at the cutoff for mild depressive symptoms (The MacArthur Foundation Initiative on Depression and Primary Care, 2004).

Data Analysis

The card sorting task was analyzed via quantitative data; each card was assigned a number and then the mean rank for each card was determined for each participant. Consistent with past card sorting methodology, cluster analyses, a commonly used statistical method for grouping complex data, were conducted to analyze the card sorting data (Anderberg, 1973; Usability.gov). For both the face-to-face and the app barrier cards, a hierarchical cluster analysis was conducted to determine the number of clusters appearing in the data set. This number was used to then conduct K-means cluster analysis to determine membership of cards within the different clusters. These analyses were conducted for the ranked means of the cards for both groups, as well as for the ranked means with the standard deviations for both card sets. Two analyses were conducted to: 1) identify the most important barriers (ranked means only to provide an indication of the average ranking of barriers); and 2) how consistently barriers were ranked as important (ranked means and standard deviations to provide an indication in the variance of ranked barriers).

Results

Participants

Table 2 displays the sample characteristics for the card sorting tasks. While equal numbers of participants above and below the threshold for a referral for therapy were anticipated, one extra person below the threshold was enrolled. Thus, nine participants were above the threshold for a referral (PHQ-9 $\geq$ 10) and 11 were below the threshold for a referral (PHQ-9 <

10). The sample was comprised primarily of females (75%) and non-Hispanic Caucasians (65%), with a mean age of 37.2 (Standard Deviation = 12.2). Those meeting criteria for a referral to psychotherapy had significantly higher depressive symptom severity (14.4 vs. 3.8, $p < .001$) and a significantly higher prevalence of past depressive episode(s) (77.8% vs. 18.2%, $p = .008$).

Face-to-Face Delivery Barriers

Hierarchical cluster analysis indicated four clusters for the face-to-face barrier task. Table 3 displays the four groups, as determined via K-means cluster analyses. The groups are listed in order of strength of the barrier, with Group 1 being the greatest barriers and Group 4 being the smallest barriers. Variance represents the clusters created by mean ranks only and Consistency represents the clusters created by including both the mean ranks and standard deviations. Differences between the rows therefore indicate variance in how highly a certain barrier was ranked. Cost was identified as the single most important barrier to face-to-face treatment. Cost was consistently followed by lack of insurance coverage and motivation, stigma, concerns about effectiveness and being seen while emotional, time for session travel and attendance, and talking with someone unknown about private topics. Barriers identified as being smaller or not as cumbersome (i.e., childcare, distance, etc. in Groups 3 and 4) were identified less consistently, as evidenced by discrepancies between the Variance and Consistency analyses. While all of the barriers included are consistent with past descriptions of barriers to face-to-face treatment for adults with depression (Mohr et al., 2006), the importance of some barriers appears to have decreased in the current evaluation (i.e., those included in Groups 3 and 4).

App Delivery Barriers

Hierarchical cluster analysis indicated four clusters for the mobile barrier task. Table 4 displays the four groups, as determined via K-means cluster analyses. Similar to Table 3, the

groups are listed by strength of the barriers, with Group 1 being the greatest barriers and Group 4 being the smallest barriers. Concerns about effectiveness, data access and privacy, cost of data package, bugs in the system, availability of Wifi, and misfit of features to needs were consistently rated as the top barriers to mobile treatments. Greater discrepancies occurred in the next highest groupings of barriers, however, concerns over not receiving enough feedback and lack of guidance were the next greatest barriers, on average.

Discussion

The present study identified user perceived barriers to face-to-face and app delivery of depression interventions via two card sorting tasks. Cost was consistently rated as the top barrier to face-to-face delivery, and top app barriers included concerns over intervention efficacy, app functioning, privacy, cost, and lack of guidance and tailored feedback. The common top barrier between the two delivery methods was cost, suggesting that this is a cumbersome barrier for users with depression, regardless of delivery mechanism. Examination of these barriers indicates specific design recommendations to address user concerns.

Cost was identified as a top barrier for both delivery mechanisms, but it is unclear if the same meaning was associated with both mechanisms. Cost of therapy (i.e., cost of service) has previously been detailed as a primary barrier to initiation and maintenance of face-to-face delivered treatment (Mohr et al., 2006). Ancillary costs, such as paying for transportation and childcare have also been noted (Mohr et al., 2006). Qualitative feedback indicated that participants generally interpreted “cost” as meaning the cost of service for face-to-face therapy. In apps, cost of apps (i.e., cost of service) has previously been suggested as an inhibiting factor in adaptation of mobile technologies in community health settings and across general health app consumers (Glick, Druss, Pina, Lally, & Conde, 2015; Krebs & Duncan, 2015). Cost of apps has

also been cited as a top user criticism in app user reviews (Fu et al., 2013). However, participants identified the cost of data package (i.e., ancillary costs) as a primary barrier. This suggests that ancillary costs, which are possibly hidden or unclear to a user, are of greater concern than the cost of service. This shift in concern over cost is a difference between face-to-face and app delivery of interventions for depression. As apps are being designed and disseminated with an aim to overcome barriers to traditional intervention delivery mechanisms, overlap in barriers with face-to-face interventions are particularly problematic. Cost appears to be a consistent concern across delivery mechanisms, however the focus appears to shift towards ancillary costs as opposed to service costs.

After cost, barriers to apps are related to user uncertainties around use of them as a delivery mechanism, such as data access and privacy, app functioning, guidance, and efficacy. These findings are not surprising, given previous reports indicating that information about app privacy and efficacy are frequently not communicated to users. Indeed, the majority of privacy policies for currently available apps are missing, not focused on the app itself, or require college-level literacy for comprehension (Sunyaev, Dehling, Taylor, & Mandl, 2015). Additionally, a majority of health apps have been found to pose threat to security and privacy of user data (Dehling, Gao, Schneider, & Sunyaev, 2015; He, Naveed, Gunter, & Nahrstedt, 2014). While current users of health apps generally report trust in their accuracy (Krebs & Duncan, 2015), efficacy related to cultural and symptom-specific factors have also been cited as potential barriers or concerns about smartphone intervention uptake (Derbyshire & Dancey, 2013; Genz et al., 2015; James & Harville, 2015). Further, app functionality issues, such as errors and app crashes, have previously been identified as primary criticisms from app users (Fu et al., 2013;

Khalid, Shihab, Nagappan, & Hassan, 2014). The barriers identified through card sorting are consistent with previously raised issues and concerns from app users.

Among other barriers identified for apps, was the potential user's beliefs of lack of guidance and feedback. This issue may overlap with a less primary barrier identified via the card sorting task: lack of human interaction. Integration of human support in health interventions delivered via technology has been recommended, and included in apps and other technologies, for the purposes of improving adherence, communication with care teams, and improving quality of tool use (Årsand et al., 2012; Possemato et al., 2016; Schueller, Tomasino, & Mohr, In Press). However, the majority of currently available apps for depression do not include connection to human support, nor provide personalized guidance or feedback (Shen et al., 2015). These findings highlight implications for design changes and improvements that better align with the needs and concerns of users.

Implications for Design

Table 5 details implications for design based upon identified barriers to use and uptake of apps for depression. Implications and their rationale are detailed below.

Cost. The barrier of cost is associated with the card: Cost of data package. With cost identified as a primary barrier in both face-to-face and app delivery, the design and marketing of apps for depression would likely benefit from transparency of possible costs, and an emphasis on avoiding hidden costs. While there is sometimes a cost associated with purchase of an app, participants indicated through qualitative feedback that their concern over cost is specific to the cost accrued through an app's use of their data packages. This concern leads to two recommendations. First, users should be provided a choice of whether an app will utilize wireless data, or only use data when connected to a Wifi source. Second, users should be provided clear

information at download on whether an app requires an Internet or data connection, and at what amount and frequency. The majority of apps designed for the most prevalent health conditions do not require an internet or data connection for use, following download (Martinez-Perez, de la Torre-Diez, & Lopez-Coronado, 2013); users may therefore be making assumptions about the cost of apps due to data usage beliefs that are incorrect. Further research is needed to expand cost-effective means for use of apps to deliver depression interventions and how to transparently detail all costs and data requirements of these apps to users.

Privacy and Security. The barriers of privacy and security are associated with the cards: Unsure who has access to data, and Privacy. Strategies and recommendations have previously been proposed to combat the critical issues of privacy and data safety in app design, including: data encryption, user access controls, privacy notices, and creating privacy profiles (Dehling et al., 2015; Lin, Liu, Sadeh, & Hong, 2014; Martinez-Perez, de la Torre-Diez, & Lopez-Coronado, 2015; Silva, Rodrigues, Canelo, Lopes, & Zhou, 2013; Yang & Silverman, 2014). However, to address user concerns in design through impacting user knowledge and awareness of data security and privacy, a clear and concise privacy statement at launch is recommended. Further, it is recommended that if the app accesses data from features on the phone or other apps, that this be stated explicitly. Users are more likely to view an app's access of private information as appropriate and acceptable if it fits their expectations of the app's function (i.e., a mapping app accessing current location via the GPS feature on the phone; Lin et al., 2012). Therefore, at initial launch, apps delivering depression interventions should initiate a pop-up request for access to any possible features or data collected from the phone. Links to additional information should be provided to clearly and concisely detail: 1) why this access is needed, 2) if and how the app functionality will be impacted if this access is not allowed, and 3) the storage and confidentiality

of retrieved data from these features. These permissions should also be editable over time, in case selected access permissions are inconsistent with later user interactions and needs of the app. Future research is required to understand the impact of these design recommendations on improving user comprehension and sense of control over app privacy and security.

Efficacy and Functionality. The barriers of efficacy and functionality are associated with the cards: Concerns about effectiveness, Misfit of features to needs, Bugs in the system, and Wifi access. Users expressed concern over an app's abilities to meet the treatment needs of depression, and to function with limited error (i.e., crashing, bugs in the system). To meet the concern over efficacy for a user's symptoms, video testimonials, featuring demographically representative personas, are recommended for apps delivering interventions for depression. Information delivered via Internet browsers has been found to be believed as specifically targeting a user and to be rated more favorably with the inclusion of video testimonials. This belief is strengthened, even when compared to similar testimonials presented via text or picture (Appiah, 2006). Further, video testimonials with demographically representative personas have been noted as user requirements for other types of health apps (Gilliam, Martins, Bartlett, Mistretta, & Holl, 2014; Schnall et al., 2015). User satisfaction with video testimonials may be evaluated before deployment via usability testing. Further, usability testing and quality assurance evaluations should be employed before releasing apps for depression, in an effort to identify and remove likelihood of app crashes and bugs (Khalid et al., 2014; Nielsen, 1993). In addition, an easily located help option should be made available on apps delivering depression interventions, so users have the ability to troubleshoot should app functioning become problematic. The help button should link with frequently asked questions (FAQs), as well as an option to connect with

live support. Future evaluations of video testimonials and troubleshooting efforts are necessary to identify how concerns over efficacy and functionality are impacted by these changes.

Feedback, Guidance, and Human Interaction. The barriers of efficacy and functionality are associated with the cards: Not enough feedback, Concerns over lack of guidance, and Lack of human interaction. Given benefits identified in web-based delivery platforms, providing concurrent human support, such as coaching via phone, text, or messaging, has emerged as a possible solution to concerns over lack of feedback, guidance, and human interaction in apps delivering depression interventions (Mohr, Burns, et al., 2013; Schueller et al., In Press). Coaching has been identified as a means to enhance supportive accountability, a construct that is intended to increase adherence, which may impact outcomes in use (Mohr, Cuijpers, & Lehman, 2011). Indeed, coached interventions have demonstrated significantly better adherence than non-coached interventions for depression (Cuijpers, Berking, et al., 2013). Rather than provide therapeutic interventions, coaches are intended to increase engagement and motivation with a technology-delivered intervention by reinforcing successful use (Berger, Hammerli, Gubser, Andersson, & Caspar, 2011; Mohr, Duffecy, et al., 2013). Aiding users in full and confident engagement with an app may address the issue of lack of guidance. A future design option may include algorithms that initiate feedback based on specific user behaviors or detected user contexts (Burns et al., 2011; Saeb et al., 2015). However, more research is needed to understand the best means to implement interventions related to passive behavior detection in those with depression. An increase in the use of human support via coaching will need to be evaluated for its impact on user perceptions of feedback, guidance, and human interaction while using apps for the delivery of depression interventions.

Limitations

Limitations and caveats should be considered in the interpretation of these findings. First, while the sample size was sufficient for a card sorting task (Nielsen, 2004), the sample was comprised of urban and primarily younger, non-Hispanic Caucasian users. It is unclear how well these findings extend to users in differing geographical locations and demographic groups. However, the process of identifying barriers with participants in an in-person setting was established as feasible. Second, the sample was a mixed group of those with no depressive symptoms to severe depression, with the majority in the mild symptom range. It is unclear if similar groupings of barriers would be identified with a more severely depressed sample, or those with comorbid psychiatric or health conditions. Despite concerns of generalizability to more severe samples, this sample represents the diversity of symptoms experienced across the typically relapsing and remitting course of depression (Judd et al., 1999; Mueller et al., 1999; Paykel, 2008). Third, it is possible that the participants inferred different meanings for the barriers listed on the cards. For example, the cards "Unsure who has access to data" and "Privacy" were typically ranked differently despite having similar meanings. While qualitative feedback was utilized to better understand rankings and groupings of the cards, future research utilizing card sorting to identify barriers would benefit from uniform definitions for each card.

Conclusions

To the best of our knowledge, this is the first use of a card sorting task to identify user-perceived barriers to apps. Smartphones stand as a promising delivery mechanism for overcoming barriers to traditional delivery of depression interventions. However, cost remains a consistent barrier across face-to-face and app delivery. Other barriers to the use of apps for the delivery of depression interventions relate to uncertainties around apps as a delivery mechanism.

Implications for design to address these barriers include: limiting wireless data usage, clearly stating possible costs and privacy/access options at download, including demographically-representative video testimonials, conducting usability testing and quality assurance evaluations, and including human support. Future research should evaluate the impact of changes in design and marketing of apps delivering interventions on perceptions of barriers for users with depression.

Chapter III: Exploring User Learnability and Learning Performance on an App for Depression

Introduction

Roughly two-thirds of Americans own smartphones, and nearly 20% of all Americans rely on this technology as their only method for internet access (Smith, 2015). This tremendous growth in smartphone access and use has made it an attractive avenue for the delivery of behavioral health interventions via apps. As of 2015, most apps with a focus on mental health were designed with a narrow or single functionality, such as providing information to users with a way to enhance learning about their mental health symptoms, or their management (IMS Institute for Healthcare Informatics, 2015; Shen et al., 2015). One such narrow functionality app is Thought Challenger, an app designed to promote the use of an intervention skill for depression (Lattie et al., In Press). This paper will explore the role of learning in narrow functionality apps for depression through usability testing of a specific app, Thought Challenger.

Thought Challenger

Thought Challenger is one app in the IntelliCare suite, a collection of apps in which each app focuses on one behavioral strategy commonly utilized in the treatment of depression or anxiety (Lattie et al., In Press). Thought Challenger instantiates thought restructuring, the core strategy in cognitive therapy (CT), that involves identifying and appraising maladaptive thoughts and creating adaptive counter thoughts (J. S. Beck, 2011; Lattie et al., In Press). CT was therefore utilized as the theoretical framework to guide the design of Thought Challenger. Thought Challenger is intended to teach users a specific skill, thought restructuring, through repeated interactions with the app.

Thought Challenger is narrow in its functionality, with a tool and review function. Figure 1 displays screen shots of the tool interactions and review function of Thought Challenger. The tool interaction of Thought Challenger involves five steps to engage a user in thought restructuring: 1) "Catch It," wherein users enter a recent maladaptive thought; 2) "Check It," in which users are asked reflective questions regarding their thought; 3) "Choose a Distortion," which asks users to identify in which type of cognitive distortion their thought likely falls; 4) Consider reflective questions tailored to the chosen type of distortion; and 5) "Change It," wherein users enter in a more adaptive thought. Within steps one and five, Thought Challenger provides examples of possible maladaptive and adaptive thoughts, which users may select and utilize in their interaction with the thought restructuring tool. Thought Challenger also provides a review function so users can see entries of all thoughts, listed by automatic thought, rational response, distortion, and date and time of interaction. Through its narrow functionality, Thought Challenger is a skill-based app attempting to promote user learning around the CT intervention strategy of thought restructuring.

Learning

Learning in Apps. Merriam-Webster defines learning as "the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something" (Merriam-Webster.com); however, in the context of apps, learning can be defined in many ways. Interventions delivered via apps are often formed through the collaborative process of multi-disciplinary teams, including, but not limited to, the fields of psychology, public health, engineering, and design (Schueller, Munoz, & Mohr, 2013). Therefore, posing the question of what and how users should learn from an app for depression may receive drastically different responses, based upon different theories and models created to achieve the goals of "learning." A

framework is therefore useful in classifying learning objectives, assessment, and outcomes. The present study utilized the revised Bloom's Taxonomy to define learning from a depression app as: a user experiencing an increase or change in a type of knowledge (i.e., factual, conceptual, procedural, or metacognitive) through the execution of a specific action (i.e., remember, understand, apply, analyze, evaluate, or create), facilitated by an app (Anderson, Krathwohl, & Bloom, 2001; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Krathwohl, 2002).

Learning in Cognitive Therapy as a Framework for Learning in Thought Challenger. CT focuses on educating patients about the impact of their thoughts on their mood, while providing insight into opportunities for change and ultimate symptom reduction (J. S. Beck, 2011). Patient learning and application of skills are noted to be among the possible mechanisms supporting symptom change in cognitive interventions (Barber & DeRubeis, 1989; Hundt et al., 2013). Thought Challenger was designed to promote the learning and application of skills associated with symptom change in CT. However, the effectiveness of Thought Challenger in achieving this design aim is unknown. To evaluate user learning associated with using CT as a theoretical framework, the goals of learning for Thought Challenger must be identified and examined.

Goals of Learning in Thought Challenger. The goals of user learning from Thought Challenger are defined via the following objectives, based in specific actions needed to engage in thought restructuring using the app:

1. Identify specific thoughts that are maladaptive, whether from recent memory or through recognition of similar thought in examples
2. Understand different types of thought distortions through provided definitions
3. Classify maladaptive thoughts into distortion categories

4. Generate specific thoughts that are adaptive, whether from using reflective questions or through identification with an example thought
5. Identify common thought patterns through review of entries

Classifying Learning Objectives

Revised Bloom's Taxonomy. The revised Bloom's Taxonomy is used to ground the learning objectives of Thought Challenger into a generalizable framework (Anderson et al., 2001; Krathwohl, 2002). The original Taxonomy provided a framework for classifying statements identifying what is anticipated to be learned for a given product (Bloom et al., 1956). The revised Taxonomy expands this framework into two dimensions: Knowledge and Cognitive Processes. Additionally, categories were expanded to include concepts not well identified at the time of the original Taxonomy's development, such as metacognitive knowledge (i.e., thinking about one's thinking process, including how and when to use a strategy for a given problem; Anderson et al., 2001; Krathwohl, 2002; Pintrich, 2002). While many psychological theories may be used as theoretical frameworks for the design of mental health apps (e.g., cognitive, behavioral, social learning theory); the revised Taxonomy provides a generalizable classification system to guide evaluation of learning goals in these apps, regardless of differences in psychological approach. Indeed, Bloom originally intended the Taxonomy as a means to create a common language across disciplines and subject matters to evaluate educational objectives and ultimate outcomes (Bloom et al., 1956).

Table 6 displays the Knowledge and Cognitive Process Dimensions of the revised Taxonomy, including the attributions of each level (Anderson et al., 2001; Krathwohl, 2002). Learning objectives map onto the revised Taxonomy by considering the phrasing of the objectives, as they are generally framed by 1) some content and 2) a description of desired user

actions with that content. Learning objectives therefore typically contain a noun or noun phrase (i.e., the content) and a verb or verb phrase (i.e., the desired actions). The revised Taxonomy's dimensions map onto learning objectives such that the Knowledge Dimension is applied to the content (noun) and the Cognitive Process Dimension is applied to the desired actions (verb; Krathwohl, 2002). For example, for the learning objective, "the user will be able to *recall* (verb) a *positive thought* (noun)," the Knowledge Dimension Attribute of factual knowledge (i.e., specific details: a positive thought) and the Cognitive Process Dimension Attribute of remember (i.e., recall) are applied. The revised Taxonomy can classify learning objectives via the desired action and subject matter.

Classifying Thought Challenger. Table 7 displays the revised Taxonomy Table (Krathwohl, 2002), which shows the overlap of objectives across the two revised Taxonomy Dimensions. The first learning objective of Thought Challenger, "*Identify specific thoughts* that are maladaptive, whether from recent memory or through *recognition* of *similar thought* in examples" applies to the Knowledge Dimension Attribute of factual knowledge (i.e., nouns: specific thoughts, similar thought) and Cognitive Process Dimension of remember (i.e., verbs: identify, recognition). The first learning objective therefore falls in 1A (Remember/Factual Knowledge) of the Taxonomy Table. The second objective, "*Understand different types of thought distortions* through provided definitions," falls in 2A of the Taxonomy Table, requiring both factual knowledge (i.e., different types of thought distortions) and the Cognitive Process of understand (i.e., understand via interpreting distortion definitions). The third objective falls in 2B, requiring both conceptual knowledge (i.e., distortion categories) and the Cognitive Process Attribute of understand (i.e., classify), as users must "*Classify* maladaptive thoughts into *distortion categories*." The fourth learning objective, "*Generate specific thoughts* that are

adaptive, whether from *using reflective questions* or through *identification* with an *example thought*," is identified as a complicated objective once mapped onto the revised Taxonomy. Indeed, it contains three noun and verb groupings. The fourth learning objective therefore falls in 6D, 3D, and 4C on the table, requiring procedural (i.e., example thought) and metacognitive knowledge (i.e., specific thoughts that are newly generated, reflective questions) and the Cognitive Processes of apply (i.e., using), analyze (i.e., identify or differentiate), and create (i.e., generate). Finally, the fifth learning goal, "*Distinguish common thought patterns* through *review of entries*," also falls in multiple places on the table. The Knowledge Dimension Attributes of conceptual (i.e., common thought patterns) and factual knowledge (i.e., entries) and the Cognitive Process Attributes of analyze (i.e., distinguish) and evaluate (i.e., review) are indicated. Therefore, the final learning objective falls in both 4B and 5A on the Taxonomy Table (Krathwohl, 2002).

Evaluating Learning Objectives

With the objectives and framework for understanding learning for Thought Challenger defined and classified using the revised Taxonomy, user learning from interacting with Thought Challenger (i.e., how well the objectives are achieved) can be systematically examined via usability testing methodologies. Usability testing is a method of evaluation that involves testing users' interactions with a product and system to improve design. This process is intended to ensure that a technology is intuitive and easy to use. Sometimes confused with informal inquiry of user opinions of a product, usability testing requires systematic observation of a planned task or scenario carried out by an actual or potential user (Usability.gov). Nielsen identified five primary attributes of usability: 1) Learnability (i.e., the ease with which a user can accomplish tasks upon initial encounter); 2) Efficiency (how accurately and completely a user can complete

tasks in a given time); 3) Memorability (how easily and proficiently a user can complete tasks after a delay in use); 4) Errors (how frequently a user makes and how easily a user can recover); and 5) Satisfaction (how pleasant a user finds a product; Nielsen, 1993). The International Standards Organization (ISO) provides standards for usability testing which define how to identify the information necessary for a designer to take into account when specifying or evaluating usability of an evaluated product (Tullis & Albert, 2008). These techniques are used in engineering and computer science to evaluate and refine products, and are being used with increasing frequency in the context of behavioral health interventions delivered via technologies (e.g., Ben-Zeev et al., 2013; Mohr et al., 2015).

Purpose

The purpose of this study is to understand learning in the context of an app for depression, Thought Challenger, via usability testing methodologies. The current study will test three questions to evaluate the efficacy of the app in achieving the intended learning objectives: 1) How well does a user initially interact with the Thought Challenger app without instruction; 2) Is a user is able to learn the skill of cognitive restructuring from the app; and 3) Does use of Thought Challenger change baseline knowledge of cognitive therapy elements.

Methods

The design and execution of the learning evaluation of Thought Challenger followed a framework proposed for evaluating apps (Zhang & Adipat, 2005). Figure 2 displays the use of the framework to design the present study, and methodological decisions are explained below.

Identifying Research Questions and Objectives

The framework begins with the identification of research questions and locating appropriate objectives with which to answer these questions. The purpose of the present study

was to evaluate how well a user will learn a depression intervention skill through the use of an app, without first reviewing any instructions. The evaluation of learning without instruction is important, given the phenomenon of the Paradox of the Active User. This paradox refers to known patterns of use indicating that despite the likely benefits of reading instructions, users are unlikely to engage with instructions or help materials prior to use (Carroll & Rosson, 1987). The Paradox of the Active User has been found to extend to the use of apps (Koole, 2009). Therefore, apps should be able to achieve their aims through intuitive design (Bedford, 2014). Consequently, the objectives of the testing were to: 1) Identify how a user interacts with the Thought Challenger thought restructuring tool without instruction or didactic material; 2) Examine if a user learns to use the app within an acceptable time limit and with a low error rate; and 3) Measure change in knowledge of cognitive therapy intervention elements following initial use of Thought Challenger.

Testing Method

Testing methods can employ laboratory experiments or field studies (Zhang & Adipat, 2005). Laboratory testing was chosen for the evaluation for several reasons relating to increasing internal validity of the findings. While remote moderated or unmoderated testing allows for more external validity (i.e., generalizability to contextual situations) and lower cost, laboratory testing provides more benefits for the current evaluation. Data quality is anticipated to be greater, as well as providing the opportunity for increased qualitative insights, as the facilitator can probe issues through observation of interface and user reactions. Laboratory testing also allows for excellent metric quality (Sauro, 2012a). A limitation to laboratory testing is being unable to evaluate possible problems related to contextual factors in real life environments, such as

connectivity problems. However, Thought Challenger is a native app and this possible limitation is therefore not applicable for the present evaluation.

Tools Used

The framework proposes a choice in tools used for the testing of mobile apps, ranging from emulators or paper prototypes to actual mobile devices (Zhang & Adipat, 2005). Emulators are commonly used in early evaluation to support design decisions. However, in this case, Thought Challenger is a fully functioning app that has been deployed on the Google Play Store (Lattie et al., In Press). Further, testing Thought Challenger on a mobile device facilitates the likelihood of an accurate evaluation for the app, as it portrays the actual physical constraints and characteristics that users experience in actual use (Harrison, Flood, & Duce, 2013; Zhang & Adipat, 2005).

Selecting What to Measure

Informed by the research questions and objectives, the framework next requires selecting usability attributes (Zhang & Adipat, 2005). Attributes are usability features that enable measurement of different usability qualities of technology products (Nielsen, 1993). Based upon the objectives identified earlier in the framework, the research questions will be addressed using the attributes of usability testing associated with learnability and learning performance. Table 8 details the attributes selected, how they were measured, and how they apply to the objectives of the usability testing and learning in Thought Challenger.

Learnability. Learnability is defined as the level of ease through which a user gains proficiency with an app (Harrison et al., 2013). Learnability has been measured in usability testing by measuring the time to complete a task at initial use and the amount of errors and speed through initial uses (Harrison et al., 2013; Kiili, 2002; Parush & Yuviler-Gavish, 2004; Ziefle,

2002). Learnability was measured in this testing as a means of evaluating learning objectives one and two: identify how a user interacts with the Thought Challenger thought restructuring tool without instruction or didactic material and examine if a user learns to use the app within an acceptable time limit and with a low error rate.

Learnability of the Thought Challenger tool was ascertained through multiple methods. First, time to completion for unguided interactions with the tool was measured across two separate attempts. As users report spending about five minutes or less to learn how to use an app (Flood, Harrison, Iacob, & Duce, 2012), successful time to completion was defined as an interaction completion time of five minutes or less. Second, learnability was measured by error rate. Errors were categorized as slips (i.e., an unintended action with the correct goal, such as a typo), mistakes (i.e., a behavior with an incorrect goal, such as typing in today's date rather than a date of birth), or fatal errors (i.e., an error that prevents the user from completing the task even with provided instruction/guidance; Norman, 2002; Sauro, 2012b). Error rates were obtained by dividing the total number of errors made by the number of error opportunities. Error opportunities are the total number of actions a user must complete to finish an interaction without errors (Sauro & Kindlund, 2005). For the purposes of the structured interaction with Thought Challenger, the number of error opportunities was 21. To the best of our knowledge, the literature does not define an ideal error rate for initial app use. Therefore, an error rate will be established and any violated usability heuristics will be identified. Third, learnability will also be measured via ability to appropriately complete a thought restructuring exercise using the Thought Challenger app. As thought restructuring can be a difficult skill for patients to grasp on initial attempts (J. S. Beck, 2011; Gumport et al., 2015; Rees, McEvoy, & Nathan, 2005), a successful rate for this measure of learnability will be that 63% or more of entries into the app

are judged as correct examples of thought restructuring. This rate is based upon findings of patient abilities to complete thought records on their own during face-to-face delivery of cognitive interventions (Rees et al., 2005).

Learning Performance. Learning performance is an attribute of usability relating to actual impact of a technology on performance of a task or acquisition of knowledge, such as the ability of a technology to aid in increasing capabilities to complete assignments in a classroom (Luchini, Quintana, & Soloway, 2003). As the testing of this study occurred during single, in-lab sessions, learning performance was measured via scores on a pre/post test of cognitive therapy knowledge and skills. Successful learning performance was defined in this study as a significant increase in the score of a questionnaire evaluating cognitive therapy knowledge and skills in a pre/post test administration. Learning performance was measured in this testing as a means of evaluating objective three: measure change in knowledge of cognitive therapy intervention elements following initial use of Thought Challenger.

Data Collection Approaches

The final decision point in the framework is determining data collection approaches. The decision to conduct laboratory testing provided guidance to use more traditional methods, such as observation of a standardized protocol, use of questionnaires and interviews, and review of data logs entered into the mobile device utilized for testing. Traditional data collection methodologies have been used in other evaluations of apps successfully (Harrison et al., 2013; Mohr et al., 2015; Zhang & Adipat, 2005). Further, new data collection methodologies are often related to addressing the challenges of conducting usability testing via field studies and are therefore not necessary for the present study. Hence, traditional usability data collection approaches were chosen for the testing of Thought Challenger. Specifically, data collection

included: 1) video/audio recording of the interactions; 2) standardized interview questions, with the option to prompt regarding specific behaviors or observations; 3) questionnaires (see Measures section); 4) timing of all interactions via stop watch; and 5) recording of all user actions into the app's thought restructuring tool (i.e., entry of thought, assignment of type of thought distortion, etc.).

Procedure

Recruitment of participants occurred from July to August 2015 from online postings in the Chicago area, resulting in the participation of 20 adults. Inclusion criteria required that participants were at least 18 years of age, able to attend an in-lab testing session, and able to speak and read in English. As depression is a condition that is frequently chronic, characterized by patterns of remissions and relapses (Judd et al., 1999; Mueller et al., 1999; Paykel, 2008), equal numbers of participants currently above and below the criteria for a referral for psychotherapy were recruited (The MacArthur Foundation Initiative on Depression and Primary Care, 2004). This sampling ensured that learning objectives were being measured with likely end users, ranging from those with no or mild depressive symptoms (subthreshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score less than 10) to those with moderate or severe depressive symptoms (threshold for a referral to psychotherapy, as measured by the Patient Health Questionnaire-9 score greater than or equal to 10; Kroenke & Spitzer, 2002). Participants who completed in-lab usability sessions were compensated $20 in petty cash for their time and participation. In compliance with the University's Institutional Review Board (IRB), participants completed an online screening consent prior to the collection of any data and were consented in-person for the usability testing session.

Prior to the testing of Thought Challenger, participants engaged in a card sorting task to identify barriers to the use of apps for depression (see Chapter II). Following this, participants were provided the prompt: "Thought Challenger helps you gain control of how you feel and what you do by teaching you to notice and challenge negative and unhelpful thoughts. Thought Challenger is built on cognitive therapy - a structure that has been found in clinical studies to be useful in examining negative thoughts and reframing them to help you feel better and do the things you want to do." Users were then instructed to pick up the Android phone used for testing (laying on table directly in front of user), open the Thought Challenger app, challenge a recent negative thought, and inform the testing moderator when the user believed the task was completed (i.e., Thought Challenger Learning Objectives 1-4). The interaction was timed and recorded, noting errors and alternative paths made in completing the interaction. Users were then queried about any alternative paths taken to complete the interaction, if they were able to find the log of the tool interaction they just completed (i.e., an element of Thought Challenger Learning Objective 5), and if they were able to find more information about the app (i.e., Frequently Asked Questions or Help sections). These interactions were also recorded and timed, and allowed for a delay between the two tool interactions measured. Once completed, the users were prompted: "Now, please log another recent negative or unhelpful thought you have had." This interaction was also timed and observed, and all entries into the tool were recorded for later review (i.e., Thought Challenger Learning Objectives 1-4). Following a brief interview of the user impressions of Thought Challenger, users completed questionnaires on a lab computer.

Measures

Study data were collected and managed using REDCap electronic data capture tools hosted at Northwestern University (Harris et al., 2009). REDCap (Research Electronic Data

Capture) is a secure, web-based application designed to support data capture for research studies, providing 1) an intuitive interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures for seamless data downloads to common statistical packages; and 4) procedures for importing data from external sources.

At screening, participants were asked to provide demographic information (i.e., gender, race/ethnicity, age, education, and employment status). Further, they completed the Patient Health Questionnaire-9 (PHQ-9) and Cognitive Therapy (CT) Tool Knowledge and Skill Pre-Test at screening (Kroenke & Spitzer, 2002; Wright et al., 2002). Following completion of the interactions with Thought Challenger in the usability testing session, participants completed the CT Tool Knowledge and Skill Post-Test, which is identical to the Pre-Test.

The PHQ-9 is a 9-item self-report instrument measuring depressive symptomology with scores ranging from 0-27 (Kroenke & Spitzer, 2002). The CT Tool Knowledge and Skill Pre/Post-Test is a measure adapted from the Cognitive Therapy Awareness Scale (CTAS; Wright et al., 2002). The CTAS is a measure evaluating understanding of cognitive therapy constructs and skills. The language in the CTAS was modified to reflect only language and concepts presented in the Thought Challenger app. The range of possible scores is 0-40. The CT Tool Knowledge and Skill Pre/Post-Test were administered at screening (pre) and after interacting with the app during the testing session (post). These time points allowed for about one week's delay between the pre- and post-test administration, with the intent of negating possible priming effects associated with pre/post-tests.

Data Analysis

The entries of user interactions were collected to measure success of users in Thought Challenger tool use, that is, identifying how effectively users engaged in thought restructuring on the app. Following the completion of all testing sessions, five doctoral-level clinical psychologists blindly rated participants' entries of maladaptive thoughts, assignment of type of cognitive distortion, and entries of alternative thoughts across their two interactions with the tool (such that each complete entry was rated by two separate psychologists). The clinical psychologists were instructed to evaluate the entries as if they were thought records, a tool typically administered via paper hand out in face-to-face CT to enable the practice of thought restructuring (J. S. Beck, 2011). When there was conflict in the psychologist ratings (each entry was rated by two psychologists), a third clinician was invited to provide consensus on the entry.

Given the small sample size and anticipated non-normal distribution (i.e., participants ranging from no depressive symptoms to severe), nonparametric tests were conducted to analyze quantitative usability testing data. Wilcoxon signed rank tests were used to analyze comparison of time to completion of the tool interaction on the first and second attempt as well as comparison of scores before and after the interaction with Thought Challenger. To ensure there were no significant differences between the participants recruited with PHQ-9 scores above and below 10, Mann-Whitney *U*-tests were performed to compare the participants on times to completion, total scores on completed measures, and demographic variables. Chi-square tests were completed to compare categorical demographic variables. All analyses were run in IBM SPSS Statistics, version 23, at the nominal 0.05 type I error rate.

Results

Participants

Table 9 displays the sample characteristics for the evaluation of Thought Challenger. One extra participant was recruited to the PHQ-9 < 10 group, making the groups roughly equal. There was no significant difference between participants above and below the criteria for a referral for psychotherapy for age, gender, or race. Those meeting criteria for a referral to psychotherapy had significantly higher depressive symptom severity (14.4 vs. 3.8, $p < .001$) and a significantly higher prevalence of past depressive episode(s) (77.8% vs. 18.2%, $p = .008$).

Learnability

Completion Time. Table 10 displays the completion times for the Thought Challenger tool interactions. For all participants, the median time to complete an initial, unguided interaction with the Thought Challenger tool was 04:05 minutes. Sixty-five percent of the sample met the criterion requiring about five minutes or less for the first interaction (Flood et al., 2012). Median time to complete the task on second attempt was significantly faster (04:05 vs. 02:34, $p = .001$).

Error Rate. Across two interactions for each participant with the Thought Challenger tool, ten errors occurred. On the first attempt at the Thought Challenger tool interaction, nine mistakes were made, relating to attempting to interact with the Thought Challenger word cloud on the home screen (i.e., clicking on the word cloud rather than a button), selecting "Review" rather than "Challenge" to interact with the tool, and persistence in the remaining tool interactions after first entering a maladaptive thought (e.g., "I entered my thought in like it said, now what?"). No slips or fatal errors occurred for any participants across the first interaction.

On the second interaction with the Thought Challenger tool, one fatal error occurred, preventing the user from completing the task even with provided instruction and guidance due to

frustration saturation (i.e., “I don’t want to start all over again and re-enter everything.”). This fatal error occurred by the user clicking “cancel” while entering data into the tool. Thought Challenger brought the user back to the Thought Challenger home screen without saving the entered data, nor prompting the user that data would be lost. This is an example of violating the usability heuristic of error prevention (Nielsen, 1993). Of note, no slips occurred during the second interactions. While participants had in-the-moment slips, such as typos, these were not maintained in the system due to the Android operating system’s algorithm to correct slips, such as auto-populating words when a suspected typo occurs during text entry.

The total error rate for all initial interactions with the Thought Challenger tool was therefore defined by 10 (errors)/[21 (error opportunities) x 2 (number of interactions) x 20 (participants)] = .012. Therefore, the error rate on initial interactions with Thought Challenger’s tool was 1.2%.

Successful Completion of Tool Records. The majority of tool entries were rated as appropriate by doctoral level psychologists, with 75% (n = 30) success in entries of a maladaptive thought, 51.3% (n = 20) success in choice of type of thought distortion, and 74.4% (n = 29) success in the entry of an adaptive thought. Consistent with face-to-face findings, the rate of success was determined to be 63% or greater (Rees et al., 2005). The ratings provided by doctoral-level clinical psychologists indicate learnability consistent with testing aims via the Thought Challenger tool.

Learning Performance

Acquisition of Skills and Knowledge. To identify learning performance of users following use of Thought Challenger, all participants completed a pre and post-test of cognitive therapy skills and knowledge. Table 11 displays the means and standard deviations of pre and

post-test scores. A Wilcoxon signed rank test indicated significant improvement in median scores for the entire sample following the use of Thought Challenger (28.5 vs. 31.0, $p = .009$). Successful learning performance was achieved for Thought Challenger, as there was a significant increase in performance on a cognitive therapy knowledge and skills questionnaire following interactions with the app.

Consistent Performance Across Symptom Severity

Given the roughly equal split in participants above and below the threshold for a referral to psychotherapy, comparisons of these groups were made to ensure no significant differences occurred in Learning or Learning Performance. No significant differences in completion times, nor in the performance on the pre and post-test of cognitive therapy skills and knowledge before or after interactions with Thought Challenger were identified between groups ($ps > .1$).

Discussion

The present study aimed to evaluate learning, based upon specific learning objectives and theory, during initial interactions with a publicly deployed app for depression (Lattie et al., In Press). Thought Challenger presents a tool for thought restructuring without didactic material; it is learnable at an acceptable time for initial use of an app (Flood et al., 2012), and is done so with a low error rate. Results also indicate that the Thought Challenger tool promotes effective execution of thought restructuring and that cognitive therapy knowledge and skills improve significantly after initial use. Ultimately, users are able to meet the learning objectives for Thought Challenger during initial use.

Thought Challenger met the evaluated learning objectives, creating entries in the tool that met the standard of accurately reflecting cognitive therapy thought records, at a rate of about 75%. This stands in similarity to an examination of thought records in patients in a face-to-face

delivered intervention, in which 63% of patients were able to accurately complete the records as between-session homework throughout treatment (Rees et al., 2005). One possible reason for the comparable performance of participants without the guidance of a therapist is that Thought Challenger provides the option of utilizing example maladaptive and adaptive thoughts. However, in the 40 tool interactions in this testing, only seven interactions employed example thoughts in the entries (~17%). While not used frequently, the example thoughts may have provided a scaffold for participants to appropriately select and enter their own maladaptive and adaptive thoughts. Initial Thought Challenger entries are comparable in accuracy to thought records completed in the course of face-to-face interventions.

Thought Challenger was able to impact learning without requiring users to read or engage with didactic content. This is in contrast to most currently available mental health apps, which focus on providing information about symptoms and/or their management (IMS Institute for Healthcare Informatics, 2015). Indeed, when psychoeducation is presented in depression apps, a static interface is predominantly used (i.e., similar to reading an e-book; Shen et al., 2015). Thought Challenger differs from this design by training users in a skill via engagement with its tool. With continued use of the tool, users participate in ongoing practicing and testing of the skill of thought restructuring. This employment of recall and retesting is consistent with evidence-based long-term learning processes (Bjork, Dunlosky, & Kornell, 2013; Roediger & Butler, 2011). Further, Thought Challenger's objectives are met through a focus on both the Knowledge Dimension (which would be expected for an app relaying psychoeducation), and the Cognitive Process Dimension of the revised Bloom's Taxonomy (Anderson et al., 2001; Krathwohl, 2002). Including cognitive processes, such as metacognition, may improve the likelihood of learning for users, as these processes have been identified as mediators of

successful learning (Desoete, 2007; Garner & Alexander, 1989; Kim, Park, & Baek, 2009). By utilizing learning objectives that map across both the Knowledge and Cognitive Process dimensions, Thought Challenger demonstrated significant improvement in user knowledge of the intended construct, despite its contrast with typical app design (i.e., presenting static didactic information).

While Thought Challenger met the criteria for learnability and learning performance established for the present study, the evaluation indicated opportunities for improvement of the app. First, a fatal error occurred in one user's interaction with the app. This error violated the usability heuristic of error prevention (Nielsen, 1993), as this error could have been prevented through the use of a warning notification with the options to either: 1) warn the user that his/her data would not be saved if s/he continues with the action; or 2) offering the option to save the data for a later interaction before exiting to the home screen. Second, mistakes that occurred could likely be minimized through the usability heuristic of help and documentation (Nielsen, 1993). In providing more guidance to users who might be confused by the options (i.e., word cloud on home screen, whether to select "Review" or "Challenge" buttons), the likelihood of mistakes could be reduced. Through the use of their established heuristics, Thought Challenger's design could be improved to minimize the possibility of user error. In reducing the opportunities for errors, the likelihood of users to interact with the app as intended improves, thereby increasing opportunities for learning via the app.

The design of Thought Challenger was informed from the theoretical perspective of CT, however the revised Bloom's Taxonomy provided a generalizable framework in which to ground the learning objectives of the app (Anderson et al., 2001; Krathwohl, 2002). Apps for depression vary in the theoretical orientation guiding their design (e.g., cognitive, behavioral, social learning

theory, etc.; Donker et al., 2013; Shen et al., 2015). However, the use of individual frameworks to classify and evaluate apps from differing theoretical orientations limits generalizability and is likely not feasible. The revised Taxonomy may serve as a theoretically agnostic framework, adaptable to learning objectives informed by differing psychological theories. Indeed, Bloom aimed for the Taxonomy to create a common language across subject matters to evaluate educational objectives and ultimate outcomes (Bloom et al., 1956). Grounding learning in an independent framework may promote generalizability across apps for depression, regardless of original theoretical orientation.

There are several limitations and caveats that should be considered in interpreting these results. First, this was an evaluation of learnability and learning performance of Thought Challenger following initial use. It is unclear how the present results apply to long-term use, knowledge, skill application, or symptom reduction. Second, this study examined Thought Challenger in the context of users with symptom severity ranging from absent to severe depression, with the majority in the mild depressive range. It is unclear how these findings extend to users with other psychiatric or medical comorbidities, or those with very severe depression. Third, while in-lab sessions were chosen over field-testing for multiple reasons, it is possible that the presence of a session moderator impacted user confidence or performance in a way that might have differed from field use. Finally, due to geographical limitations, the sample was comprised of urban and primarily younger users; it is unclear how well these findings extend to users in differing geographical locations and demographic groups.

While the present study provides insight into user learning from initial interactions with an app targeting users with depression, it also highlights directions for future research. The concept of learning may be driven by a variety of theories. Further research is needed to better

understand how users learn concepts, skills, and how to effectively complete tasks utilizing apps for depression. This study demonstrated the utility of applying the learning objectives of a CT-informed app to a generalizable framework. Continued use of the revised Taxonomy, or similarly theoretical agnostic classification systems, may promote the generalizability of future research. Second, evaluations of how learning involved in apps impacts long-term symptom management is needed. Finally, the ideal amount of information and features utilized for users with depression requires more evaluation. Elucidating the role of learning and its impact in apps targeting depression and other mental health concerns through future research will be required to increase optimization.

In demonstrating the learnability and learning performance, the present study serves as a demonstration of the benefits of grounding learning objectives in a generalizable framework. To the best of our knowledge, this is the first use of an established framework to design and implement usability testing methods for a narrow functionality app, designed to increase user learning of a discrete behavioral strategy in users with depression. Future research is needed to explore the role of learning in such apps and how learning objectives and theory may enhance user learning, particularly in users with depression. The findings from the present study suggest that users can learn to complete a therapeutic intervention skill effectively through the use of a mobile tool, without engaging in didactic content.

Chapter IV: Pilot Randomized Controlled Trial of Behavioral and Cognitive Intervention Strategies for Depression Delivered via Mobile Apps

Introduction

Depression is prevalent and causes a significant societal burden in terms of mortality, cost, and suffering (Ferrari, Charlson, et al., 2013; Ferrari, Somerville, et al., 2013). Psychological interventions for depression are effective (Cuijpers et al., 2008), but barriers to initiating or maintaining face-to-face interventions prevent many from receiving care (Gonzalez et al., 2010; Mohr et al., 2006; Mohr et al., 2010). Therefore, alternative delivery mechanisms are increasingly being explored as a means to deliver care to adults with depression (Mohr, Burns, et al., 2013).

Smartphone apps are being developed and deployed as an avenue for delivering psychological interventions to those with depression (Donker et al., 2013; Shen et al., 2015). Apps show promise as a delivery mechanism for a number of reasons. Their instantiation within a mobile device ensures that users tend to have access throughout the day. This accessibility provides opportunities for real-time monitoring, assessment, and interventions in the real-world conditions of an individual with depression (Proudfoot, 2013). Additionally, nearly two-thirds of all Americans own a smartphone (Smith, 2015), which increases the possible reach of interventions delivered via apps. This reach also extends to younger adults, minorities, and low income users (Pew Research Center, 2014; Smith, 2015); groups which are more likely to experience stigma-related or practical barriers to traditional treatment delivery (Cooper et al., 2003; Gonzalez et al., 2010; Menke & Flynn, 2009; Mohr et al., 2006; Mohr et al., 2010; Möller-Leimkühler, 2002; Wallace et al., 2006). Finally, those with mental health concerns have expressed interest in apps monitoring their symptoms (Torous et al., 2014). Apps for depression

therefore have the potential to increase the accessibility and reach of psychological interventions for depression, and are desirable to patient populations.

The majority of currently available apps for depression that provide an intervention (as opposed to solely monitoring or providing psychoeducation) are informed by behavioral and cognitive intervention strategies (Donker et al., 2013; Shen et al., 2015). Behavioral and cognitive interventions carry the strongest body of evidence in the face-to-face administration of treatments for depression (Butler et al., 2006; Cuijpers, Berking, et al., 2013; Cuijpers et al., 2007; Dobson, 1989), and are generally believed to have equivalent efficacy when delivered face-to-face (Aderka et al., 2012; Busch et al., 2006; Cuijpers et al., 2007; Dimidjian et al., 2006; Dobson et al., 2008; Hardy et al., 2005; Hunnicutt-Ferguson et al., 2012; Tang & DeRubeis, 1999; Tang et al., 2005; Vittengl et al., 2005). The noted exception in comparative efficacy being that Behavioral Activation (BA) may outperform Cognitive Therapy (CT) in severely depressed patients (Dimidjian et al., 2006). However, the efficacy of behavioral and cognitive intervention strategies cannot be assumed to remain consistent when delivered through the new medium of apps.

Complicating the unknown efficacy of behavioral and cognitive interventions when delivered via apps is the typically low frequency of app use. Even when prescribed by a healthcare provider, mental health apps have the lowest sustained rate of use after 30 days, compared to any other health and wellness apps (IMS Institute for Healthcare Informatics, 2015). Further, the modal use of apps in a publicly deployed suite of behavioral and cognitive intervention apps was one use, with the range of average launches across the group of apps being 3.10-16.98 (Lattie et al., In Press). However, the relationship between use and symptom outcome

has not been fully defined (Donker et al., 2013). Evaluating the use of apps for depression may provide insights into subsequent outcomes for this delivery mechanism.

The aim of the current study is to pilot an evaluation of the usage and efficacy of apps for depression that instantiate a behavioral or cognitive intervention skill. This study compares the usage patterns, symptom change, and possible correlations of these constructs for apps informed by behavioral or cognitive approaches, compared to a waitlist control condition.

Method

Participants

Recruitment of participants occurred from September 2015 to January 2016 from online ads posted nationally on Craigslist. Internet based advertising ensured that recruitment reflected the growing number of people who seek help through the Internet, thereby enhancing external validity.

Participants were eligible for randomization if they: 1) Had a minimum score of 10 on the Patient Health Questionnaire-9 (PHQ-9; Kroenke & Spitzer, 2002), consistent with the recommendation of referral for psychotherapy at mild depressive symptoms (The MacArthur Foundation Initiative on Depression and Primary Care, 2004). 2) Had a minimum score of 11 on the Quick Inventory of Depressive Symptoms (QIDS; Rush et al., 2003; Trivedi et al., 2004), consistent with the criterion utilized for the PHQ-9. 3) Were able to speak and read English, as the interventions were solely developed in this language. 4) Were at least 18 years of age. 5) Owned an Android phone, as early development benefits from focus on only one platform and Android comprises the largest percent of the market (Sahota, 2014). 6) Had no visual, hearing, voice, or motor impairment that would prevent completion of study and treatment procedures. 7) Were not diagnosed with a psychotic disorder, bipolar disorder, dissociative disorder, substance

or alcohol dependence, or other diagnosis for which participation in this trial was either inappropriate or dangerous. 8) Were not severely suicidal (i.e., ideation, plan, and intent). 9) Were not receiving psychotherapy. 10) Were on a stable dose of an antidepressant medication (i.e., no dose changes for 4 weeks and did not intend to change the dose) or were not currently on an antidepressant medication.

In compliance with the Northwestern University Institutional Review Board (IRB) approval, interested participants were sent a link to the digital screening consent. Subjects agreed to participate in screening by checking a "yes" box and typing in their name. They were instructed to print out the screening consent form for their records. Those found eligible following screening were invited to participate in a baseline eligibility phone assessment. Prior to the phone interview, participants were emailed a link to the detailed digital version of the study consent form. Subjects agreed to participate in the study by checking a "yes" box and typing in their name, and were again instructed to print the consent for their records. After the study consent form was signed online, and before completing the phone assessment, detailed information regarding the consent was reviewed with the participant by study staff. Participants were compensated for all completed assessments.

Treatments

Participants were randomized to either Boost Me, Thought Challenger, or waitlist control on a 1:1:1 ratio, with a block size of 4. Boost Me and Thought Challenger participants received six weeks of weekly coaching sessions, didactic content delivery (one lesson per week), and use of the app (emailed to participants via an APK attachment they downloaded directly to their phones). The waitlist control group did not receive any intervention until the passage of 10 weeks occurred, to account for both the intervention period (six weeks) and follow-up period

(four weeks). Following their completion of the final assessment, waitlist control participants were given access to both the Boost Me and Thought Challenger apps.

Boost Me. Boost Me is a native Android app (i.e., an app downloaded directly to a mobile device and designed to work specifically with the Android operating system). Boost Me was developed at Northwestern University and instantiates activity scheduling, a core strategy of Behavioral Activation (BA), which aims to increase rewarding activities and monitoring of mood in relation to behavior (Cuijpers et al., 2007; Martell et al., 2010). Figure 3 displays screenshots of some Boost Me functions. Boost Me guides the user around the concept of "boosts," which are activities the user is asked to schedule and complete. Users are also asked to predict their anticipated mood resulting from the activity during scheduling and are later prompted to rate the actual mood following the completion of the activity. This process enables a comparison between anticipated and actual mood following a rewarding behavior. Users select "Boost Me" to complete this process when they notice a drop in mood, and select "Log Boost" to log a positive activity they noted as improving their mood, if this activity had not already been scheduled. Users are able to select from a list of suggested positive activities, or to enter their own. Boost Me also provides a review function so users can see entries of past positive activities and the associated moods they had with those activities. Finally, Boost Me included a persistent notification (i.e., a small box in the Android notification tray that would remain visible throughout the day) to prompt users to reflect if they "need a boost." Weekly emailed lessons for Boost Me included: "Getting Started with Boost Me," "Monitoring Activities," "Scheduling Activities," "Different Types of Positive Activities: Pleasure and Accomplishment," "Roadblocks to Doing Positive Activities," and "Moving Forward." Participants were emailed

the app via an APK attachment; however, Boost Me is currently publicly available for Android at: https://play.google.com/store/apps/details?id=edu.northwestern.cbits.intellicare.boostme.

Thought Challenger. Thought Challenger is a native Android app developed at Northwestern University which instantiates thought restructuring, the core strategy in Cognitive Therapy (CT) that involves identifying and appraising maladaptive thoughts and creating adaptive counter thoughts (J. S. Beck, 2011; Lattie et al., In Press). Figure 1 displays screenshots of Thought Challenger functions. The tool interaction of Thought Challenger involves five steps: 1) "Catch It," wherein users enter a recent maladaptive thought; 2) "Check It," in which users are asked reflective questions regarding their thought; 3) "Choose a Distortion," which asks users to identify in which type of cognitive distortion category their thought likely falls (e.g., black and white thinking, mind reading, etc.); 4) Consider reflective questions tailored to the chosen type of distortion; and 5) "Change It," wherein users enter in a more adaptive counter thought. Within steps one and five, Thought Challenger provides examples of possible maladaptive and adaptive thoughts, which users may select and utilize in their interaction with the thought restructuring tool. Thought Challenger also provides a review function so users can see entries of all thoughts, listed by automatic thought, rational response, distortion, and date and time of interaction. Unlike Boost Me, Thought Challenger was not designed to have a persistent notification in the Android notification tray. Weekly emailed lessons for Thought Challenger included: "Getting Started with Thought Challenger," "Harmful Thoughts," "Patterns of Harmful Thoughts," "Alternative Thoughts," "Disrupting Patterns of Harmful Thinking," and "Moving Forward." Participants were emailed the app via an APK attachment; however, Thought Challenger is currently publicly available for Android at:

https://play.google.com/store/apps/details?id=edu.northwestern.cbits.intellicare.thoughtchallenger.

No major changes occurred to either app during evaluation. As minor updates to address bugs identified via quality assurance testing or public deployment were executed, participants were emailed an APK attachment that allowed the updates to occur without altering any previously user-entered data in the apps.

Coaching

Participants randomized to Boost Me or Thought Challenger received weekly coaching via phone or email. Email was utilized as a means of contact if: 1) the participant was unable to be reached via phone, or 2) the participant requested email contact for a given week. Coaching was based on the supportive accountability model, which posits that coaches increase user adherence to apps by providing accountability in the context of a supportive relationship (Mohr et al., 2011). Coaching aimed primarily at maintaining engagement with the app, and not in providing therapeutic intervention. This coaching model has been validated in a trial of web-based treatment for depression (Mohr, Duffecy, et al., 2013). The same licensed clinician provided coaching to all participants in the trial.

Assessment

Assessments occurred at baseline, week 3 (mid-treatment), week 6 (post-treatment), and week 10 (one-month follow-up). Study data were collected and managed using Research Electronic Data Capture (REDCap), electronic data capture tools hosted at Northwestern University (Harris et al., 2009). REDCap is a secure, web-based application designed to support data capture for research studies, providing 1) an intuitive interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) automated export procedures

for seamless data downloads to common statistical packages; and 4) procedures for importing data from external sources.

Because a primary justification for using apps for treatment is to overcome access barriers, and recruitment was conducted nationally, requiring participants to attend face-to-face assessments would introduce sampling biases that decrease generalizability (Mohr et al., 2006; Mohr et al., 2010). Therefore, the baseline assessment interview was conducted via telephone. The same licensed clinician conducted all phone interviews. Self-report assessments occurred at baseline, weeks 3 and 6 (mid- and end-of-treatment), and at week 10 (1-month post-treatment follow-up) via REDCap. Assessment timing was designed to minimize assessment burden. To maximize blinding, only self-report measures were administered beyond the baseline assessment.

Measures of Psychological Characteristics. The Patient Health Questionnaire-9 (PHQ-9) is a 9-item self-report instrument measuring depressive symptomology (Kroenke & Spitzer, 2002). The PHQ-9 was the primary outcome measure and was administered at baseline, and weeks 3, 6, and 10.

The Quick Inventory of Depressive Symptomology (QIDS) is a 16-item interview intended to evaluate objective, evaluator-rated symptom severity (Rush et al., 2003). To ensure that symptoms endorsed on the PHQ-9 were not transient, the QIDS was administered at baseline, allowing study staff to assess whether the depressive symptoms have been present for at least two weeks. The Mini International Neuropsychiatric Interview (MINI) is a structured diagnostic interview to diagnose Diagnostic and Statistical Manual-IV (DSM-IV) and International Statistical Classification of Diseases and Related Health Problems (ICD-10) psychiatric disorders (Sheehan et al., 1997). The MINI was administered at baseline to determine

eligibility based upon possible comorbid conditions that would make participation in the trial inappropriate.

Measure of Usability. The System Usability Scale (SUS) is a 10-item self-report instrument measuring a user's rating of a product's usability (Brooke, 1996). The SUS was administered at weeks 3 and 6 for the Boost Me and Thought Challenger participants as a secondary outcome to evaluate the usability of the apps over time.

Safety Protocol

To ensure participant safety, any participant rating higher than "1: Several Days," on item 9 of the PHQ-9 ("Thoughts that you would be better off dead, or of hurting yourself") was prompted to also answer the Beck Depression Inventory (BDI), item 9 ("Suicidal Thoughts or Wishes, 0 = I don't have any thoughts of killing myself, 1 = I have thoughts of killing myself, but I would not carry them out, 2 = I would like to kill myself, 3 = I would kill myself if I had the chance"; A. T. Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Following the completion of this question, participants received a notification that the response would be reviewed within one business day and that s/he should go to the nearest emergency department or call 911 in the case of an emergency. If a participant rated a "2" or higher on the BDI item, study staff were notified by REDCap to trigger an administration of the Columbia-Suicide Risk Assessment via telephone (Posner et al., 2011). Additionally, if any participant endorsed suicidality during the baseline phone interview (i.e., during the administration of the MINI or QIDS), the Columbia-Suicide Risk Assessment was completed immediately. Any participants with severe suicidality (i.e., ideation, plan, and intent) were excluded from the study.

Data Analysis

Baseline demographic variables (i.e., gender, age, ethnicity, race, baseline depression scores, and receiving an active dose of antidepressant medication) were compared across treatment arms using one-way Analysis of Variance (ANOVA) and chi-square tests of association. A repeated measures ANOVA was conducted to examine depressive symptoms over time and across treatment groups. *t*-Tests were conducted to compare usage and perceptions of usability across Boost Me and Thought Challenger participants. Usage was defined by the number of times the app was launched, events or thoughts were logged, and the review function was launched during the treatment period. To examine the relationship between usage and change in depressive symptoms, bivariate correlations were run between the usage variables and change in depression between baseline and end of treatment (week six). Finally, to explore the role of coaching support, bivariate correlations were run among the number of coaching sessions and duration of calls, usage variables, and change in depression between baseline and end of treatment (week 6). All analyses were run in IBM SPSS Statistics, version 23, at the nominal 0.05 type I error rate. Post hoc tests among Boost Me, Thought Challenger, and waitlist control were run using a Bonferroni correction to prevent inflation of the overall type I error rate.

Results

Participants

The flow of participants through this study is displayed in Figure 4. Baseline participant demographic and clinical characteristics are displayed in Table 12. Among the 30 participants entered into the trial, there were no significant differences in demographics across treatment groups at baseline. Three participants allocated to Thought Challenger did not receive the

intervention, and two Thought Challenger participants were lost to follow-up following the baseline assessment.

Depression Scores

Table 13 displays depression scores over time and across groups. The results of a repeated-measures ANOVA indicated that PHQ-9 scores changed significantly over time (Baseline = 16.0 ± 4.6, Week 3 = 10.2 ± 5.6, Week 6 = 7.5 ± 5.5, Week 10 = 8.9 ± 5.4; $F(3) = 31.83$, $p < .001$), and were significantly different based upon group assignment ($F(6) = 2.78$, $p = .02$). There was no evidence to suggest that the assumption of sphericity had been violated, $\chi^2(5) = 3.12$, $p = .68$.

Time. Baseline PHQ-9 scores were significantly higher than any other time point (*ps* < .001). Mid-treatment (week 3) and end of treatment (week 6) PHQ-9 scores demonstrated a decrease in severity over time, with significant differences emerging between these time points ($p = .02$). Follow-up (week 10) PHQ-9 scores showed an increase in symptoms, with no evidence to suggest differences between scores at follow-up, compared to mid and end of treatment (*ps* > .9).

Group. Post hoc analyses with a Bonferroni correction indicated significant differences in PHQ-9 scores over time between Thought Challenger and waitlist control participants ($p = .03$). There was no evidence to suggest differences in PHQ-9 scores among Boost Me participants and the other two groups (*ps* > .2).

Usage

All app usage of Boost Me and Thought Challenger is displayed in Table 14. Three Thought Challenger participants never downloaded or used the app; all Boost Me participants downloaded and used the app. Boost Me was launched significantly more than Thought

Challenger overall (97.7 vs. 33.5, $t(18) = 2.59$, $p = .02$), and during weeks two (22.0 vs. 8.8, $t(18) = 2.25$, $p = .04$), three (14.1 vs. 4.4, $t(18) = 2.19$, $p = .04$), and four (17.5 vs. 5.1, $t(18) = 2.42$, $p = .03$). No significant differences emerged in the number of events (Boost Me) and thoughts (Thought Challenger) logged overall ($p = .22$), nor across weeks ($ps > .05$). There was no evidence to suggest differences in the amount of event (Boost Me) or thought (Thought Challenger) reviews overall ($p = .45$), nor across weeks ($ps \leq 1.00$).

Usage and Depression. Total app launches, events/thoughts logged, and review launches were not significantly correlated with changes in depression scores (i.e., end of treatment scores subtracted from baseline scores), either within the entire sample (i.e., Boost Me and Thought Challenger totals) or by group (e.g., only Boost Me totals; $ps > .05$).

Usability

Mid-treatment (week 3) mean SUS scores indicated that Thought Challenger (84.10 ± 10.43) was rated significantly higher than Boost Me (70.00 ± 14.31; $t(15) = -2.29$, $p = .04$). However, at end of treatment (week 6), there was no significant difference in mean SUS scores between Thought Challenger (88.57 ± 5.56) and Boost Me (78.33 ± 15.10; $t(14) = -1.70$, $p = .11$).

Coaching

Use of phone and email contact did not significantly differ between Boost Me and Thought Challenger participants, with the exception of week 5 contact, with Boost Me having a higher percentage of phone contact (70.0% vs. 11.1%, $\chi^2(1,19) = 6.74$, $p = .009$). Boost Me participants trended towards the completion of more coaching calls (3.70 ± 2.31), that were longer in duration (1882.75 ± 1241.80 seconds), than Thought Challenger participants (*Mean* Number of Calls = 2.22 ± 2.24; *Mean* Duration of Calls = 800.50 ± 530.79 seconds); however

these differences were not significant (*ps* > .07). The length and number of coaching calls were not significantly correlated with app usage (as defined by number of app launches; *ps* > .05).

Discussion

The present study piloted an evaluation of usage and symptom change through the employment of apps as delivery mechanisms for behavioral and cognitive intervention skills. Thought Challenger demonstrated the strongest efficacy in impacting depressive symptoms, with significant differences over time compared to waitlist control participants. However, app usage was not significantly correlated with depression outcomes. Indeed, Boost Me was used significantly more than Thought Challenger, but its impact on symptoms did not significantly differ from Thought Challenger or waitlist control participants.

Thought Challenger was used less than Boost Me, but its use produced significant changes in depression compared to no intervention. Previous usability testing of Thought Challenger highlights two possible explanations for this finding. First, users generally performed effective execution of thought restructuring using the app's tool function twice. Indeed, about 75% of Thought Challenger tool interactions with evaluated users were conducted correctly, as rated by doctoral level psychologists (see Chapter III). These findings suggest that the design of Thought Challenger teaches users to correctly identify and challenge thoughts. Second, following initial use of Thought Challenger, users demonstrated significant improvements in cognitive therapy knowledge and skills, as evaluated by a pre/post-test (see Chapter III). This suggests that limited use of the app can impact knowledge and skills related to CT for depression. The success of Thought Challenger in promoting learning during initial use may ultimately improve a user's ability to internalize the skill and apply it to real-world situations in

the moment. In doing so, Thought Challenger may be able to target depressive symptoms effectively.

Boost Me demonstrated a decrease in symptoms with significantly higher use than demonstrated in Thought Challenger. Indeed, the average number of app launches of Boost Me over six weeks was nearly 100 times. This stands in contrast to much lower usage patterns of open access apps with discrete interventions for depression (Lattie et al., In Press). There are multiple possible explanations for this amount of use. First, in scheduling a "boost," users are encouraged to use the app twice: one time to schedule the activity and another time to rate how the activity actually impacted their mood. This promotion of two uses for each planned boost is in contrast to Thought Challenger's single use when restructuring a thought. However, while this design may explain some increase in use, actual use indicates that on average, slightly more than half of the scheduled activities were later completed in the app (8.20/14.70). This means that slightly less than half of the planned boosts only resulted in one use of the app. Second, Boost Me had a persistent notification prompting users to reflect whether or not they needed a "boost." This constant reminder in the notification tray may have prompted increased use. Finally, Boost Me was designed to promote positive behaviors, with the aim of providing an immediate improvement in mood. These behaviors include self-generated and auto-generated suggestions. In contrast to Thought Challenger, the design of Boost Me may promote reliance upon the app to brainstorm or select rewarding behaviors with which to engage when in a lowered mood state. This possible reliance may promote frequent and ongoing use of the app. There are multiple aspects of the design of Boost Me that may have influenced higher usage patterns than Thought Challenger, or similar publicly deployed discrete intervention apps.

Limitations of the current work should be considered in the interpretation of these findings. First, as a pilot trial, the sample size was small. Small samples are typically underpowered to identify significant effects, and can also introduce potential sampling biases. However, despite a small sample, significant differences were identified, indicating large effects occurring in this trial. Retention was also strong, with at least a 90% response rate at all assessment time points, demonstrating the feasibility of executing a larger trial evaluating apps for depression. Further, as the sample was recruited via online advertisements, it also demonstrates the feasibility of this recruitment tactic for larger trials. Second, the same person performed the roles of investigator, assessor, and coach. This increases the likelihood of investigator bias impacting the evaluation of clinical symptoms. However, all follow-up assessment time points were conducted via self-report questionnaire to eliminate this bias. Additionally, separate clinical and research doctoral-level supervisors oversaw the respective execution of study elements to ensure appropriate methodologies were employed. Third, Thought Challenger and Boost Me underwent different evaluative processes. Thought Challenger underwent quality assurance testing and public deployment before being evaluated via summative usability testing (i.e., measurement of usability on a completed product) and ultimate inclusion in this study (Lattie et al., In Press; Tullis & Albert, 2008). Summative usability testing helped to enrich the interpretations of the findings related to Thought Challenger. Boost Me underwent quality assurance testing and formative usability testing (i.e., measurement of usability on a developing product; Tullis & Albert, 2008), but was not publicly deployed prior to its evaluation. Despite this difference, Boost Me did not experience any significant bugs or malfunctions (which would likely have been identified following public deployment) and interpretations were still able to be made based on the Boost Me findings.

Fourth, those randomized to Boost Me and Thought Challenger were sent weekly lessons to support their use of the apps. As these lessons were emailed, it is unclear if and how often these lessons were read and what their impact might have been upon use and response to the apps. Finally, coaching was provided via phone and/or email throughout the intervention period, without a significant effect on usage or outcome. It is unclear if the use and response to the apps would have been impacted similarly without human support.

The present research identifies several avenues of future research. First, the findings in the current research should be replicated with larger samples. Second, as a comparative trial, it is unclear how users would benefit from exposure to both apps during the same time period. An example in face-to-face delivery of this approach is Cognitive Behavioral Therapy (CBT), in which both behavioral and cognitive approaches are employed to target depression (J. S. Beck, 2011). While CBT promotes both approaches, it allows for flexibility in which approach has the strongest emphasis at a given time, based upon symptom severity and response. Indeed, Beck and colleagues (1979) originally proposed first utilizing behavioral strategies to target depression before shifting to cognitive targets. Further, they argued that the focus should shift back to behavioral strategies, should symptoms worsen during treatment (A. T. Beck, Rush, Shaw, & Emery, 1979). This recommendation is consistent with later findings that behavioral approaches more strongly benefit severely depressed patients (Dimidjian et al., 2006). This points to another future investigation need: to evaluate differential predictors of which treatment approaches delivered via apps might be best suited for which types of users. Finally, long-term impacts and usage of these apps extending well beyond six weeks' time would provide increased insights into the engagement with and impact of apps for depression.

To the best of our knowledge, this is the first randomized controlled trial examining apps delivering a discrete behavioral or cognitive intervention strategy. The findings of the present study indicate that intervention strategies for depression delivered via apps can impact symptomology and may promote continued use over six weeks. The current study demonstrates the feasibility of future research regarding the delivery of behavioral and cognitive intervention strategies via apps for depression.

Chapter V: Conclusion

The expert panel convened by the National Institute of Mental Health and the Agency for Healthcare Research and Quality determined that theoretical and research paradigms from multiple disciplines must be refined and integrated to reach and serve those with mental health needs (Mohr, Burns, et al., 2013). The projects from the present research reflected this recommendation by executing two necessary forms of research in evaluating apps for depression: 1) usability testing (Chapters II and III) and 2) a randomized controlled trial (RCT; Chapter IV). The problems, goals, and findings of the present research are detailed in Table 15. Brief descriptions of usability testing and RCTs are detailed below, followed by implications of this research in the context of Behavioral Intervention Technologies (BITs).

General Overview of Usability Testing and RCTs

Usability Testing

Through systematic observation of a planned task or scenario carried out by an actual or potential user, usability testing is a method of evaluation that involves testing users' interactions with a product and system to improve design (Usability.gov). This process is intended to ensure that a technology is intuitive and easy to use, with the importance ranging from life saving (e.g., in use of hospital equipment), to societal impacts (e.g., the ballot counting system utilized in Florida for the 2000 presidential election), to convenience (e.g., saving time on a task; Tullis & Albert, 2008). This process also provides actionable answers to questions that are critical for organizations, researchers, clinicians, etc., which are developing or using products. Usability testing is often iterative, applying a "test, fix, test" paradigm to the development of a product. Testing is often completed with a sample of about five users evaluating each new iteration. While some believe more than five participants are needed to evaluate usability of a product for a

given iteration (Faulkner, 2003; Sauro, 2010; Woolrych & Cockton, 2001), the general consensus is that the majority of usability issues can be identified with a sample size of five (Lewis, 1994; Nielsen & Landauer, 1993; Tullis & Albert, 2008). With relatively small sample sizes for each iteration, usability testing provides a systematic evaluation of research questions that require the measurement of human behavior in an interaction with a technology or product (Tullis & Albert, 2008).

Data. Usability testing is measured using metrics. Usability metrics include the time needed to complete a task, level of satisfaction, number of errors, etc. All usability metrics must be: 1) indirectly or directly measurable, and 2) quantifiable (Tullis & Albert, 2008). While qualitative data may be gathered during the course of a usability testing session, the metrics are the primary measure of usability. Usability metrics differ from other metrics because usability metrics reveal information about an interaction between a user and a technology. How this interaction is defined is typically measured via five primary attributes of usability: 1) Learnability, the ease with which a user can accomplish tasks upon initial encounter; 2) Efficiency, how accurately and completely a user can complete tasks in a given time; 3) Memorability, how easily and proficiently a user can complete tasks after a delay in use; 4) Errors, how frequently a user makes and how easily a user can recover; and 5) Satisfaction, how pleasant a user finds a product (Nielsen, 1993). Typical consumers of usability testing data tend to be in the fields of Engineering and the Computer Sciences, with results often published via conference papers and presentations.

Strengths. Metrics resulting from usability testing provide many benefits to the development and understanding of products. First, usability testing adds structure to the design process, highlighting overall issues that could lead to costly repairs (Bevan, 2009). Indeed, it

removes the need to make design decisions from an uninformed, or "gut-feeling" perspective. Further, the metrics demonstrate if improvement is actually made as a result of design changes from one iteration to the next. Second, these data can often be collected quickly, without the need to test a product for several weeks or months (Tullis & Albert, 2008). Third, usability testing research questions specify what needs to be learned about the user experience, from perceived barriers to the use of a technology (see Chapter II), to learning from an app (see Chapter III), to satisfaction with an app across weeks of use (see Chapter IV). This specificity can lead to explicit data to address questions of usability. Finally, evaluating the usability of a product to address any notable issues increases the likelihood the product will be used (Tullis & Albert, 2008). Indeed, users are less likely to engage with a product that is difficult or perceived as a mismatch with user needs (Chiu & Eysenbach, 2010; Price et al., 2014). Usability testing provides multiple strengths to the design and use of a product.

Limitations. The potential for bias in usability testing is a limitation. Snyder (2006) argued that biases in usability testing can be minimized, but not eliminated (Snyder, 2006). Therefore, possible biases must be considered in findings. Tullis and Albert (2008) generalize usability testing biases into six categories: 1) Participants, who possess varying levels of expertise, motivation, knowledge, and comfort level in the testing setting. For example, some participants may perform a task poorly due to being observed in a testing environment; or some participants may be poorly matched to likely end users, lowering generalizability to how the technology would be used. 2) The types of tasks selected for evaluation, as some tasks may be better equipped to uncover some issues over others. This bias is of particular note for more complex technologies or products. 3) The method of evaluation, which could vary in terms of session duration, testing location, and probing participants vs. having them think aloud. For

example, asking a participant to think aloud as s/he performs a task may impact the typical workflow, as people generally do not talk aloud to themselves while working (Nielsen, 2012). 4) The artifact used for testing, as the evaluated prototype can range from a paper prototype or fully functional system. Lower fidelity prototypes have previously been found to be overrated for aesthetics, to compensate for deficiencies in this domain (i.e., noting that the aesthetics on paper prototypes are less than desirable, users will rate them higher than their true experience); and users have been found to rate a product more positively when on a more attractive device compared to a plain device (Sauer & Sonderegger, 2009). 5) The physical environment, which includes lighting, video recording equipment, or presence of a one-way mirror, also potentially impacts how a user interacts with an evaluated product. 6) The moderator(s), who like participants, may vary in experience, knowledge, and level of engagement. For example, an enthusiastic moderator may incentivize participants to attempt to perform better on a task, compared to a seemingly disinterested moderator (Tullis & Albert, 2008). These general categories must be considered when interpreting the findings of usability testing and researchers should attempt to address these biases in the design of usability testing studies.

Randomized Controlled Trials

For the purposes of the current discussion, RCTs will be considered from the perspective of behavioral medicine specialists prospectively evaluating the efficacy of a treatment intervention (Chambless & Ollendick, 2001; Friedman, Furberg, & DeMets, 1998). The evaluation must contain one or more intervention techniques (without an active intervention, the study becomes observational), as well as a control group against which the intervention technique(s) is evaluated. RCTs evaluate one primary aim, often framed as a research question (e.g., will cognitive therapy significantly reduce depressive symptomology compared to a

waitlist control condition after 6 months of treatment in adults with depression?). Secondary aims that enrich the findings of the primary research aim may also be included. The RCT is conducted with the participation of volunteers representing a subset of the population defined in the primary research aim. The assignment of participants to the intervention technique(s) or control group is determined by the formal procedure of randomization. Randomization serves to remove the potential bias of allocation, to produce comparable groups, and to ensure the validity of statistical tests of significance (Friedman et al., 1998). The number of participants required is determined based upon a predefined number, which is calculated to reflect the size needed to detect statistical significance should a true intervention effect occur. RCTs provide data regarding the comparative efficacy of a treatment intervention(s) to a control condition.

Data. RCTs strive to have data that are valid and reliable through the use of specific methods. RCTs utilize valid (i.e., the assessment measures what it purports to measure) and reliable (i.e., the assessment measures are consistent across similar people over time) assessment measures to evaluate outcomes before, during, and after the intervention period. For example, in the pilot RCT detailed in Chapter IV, the Patient Health Questionnaire-9 (PHQ-9) was used as the primary outcome measure, administered at baseline and weeks 3, 6, and 10 (Kroenke & Spitzer, 2002). The PHQ-9 has been found to be valid, such that it measures symptoms of depression; and reliable, such that it measures depressive symptoms consistently across respondents and time (Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006; Kroenke, Spitzer, & Williams, 2001; Lowe, Kroenke, Herzog, & Grafe, 2004). The diagnostic validity of the PHQ-9 has demonstrated both internal and external validity, as it measures depression effectively both in research and in primary care settings (Williams, Noël, Cordes, Ramirez, & Pignone, 2002). The PHQ-9 also yields a specific, quantitative score, enabling comparisons between and within

subjects over time. RCTs provide valid, reliable, and comparable data revealing information about a treatment's efficacy compared to a control condition. Typical consumers of these data are behavioral scientists, with results often published in peer-reviewed journals.

Strengths. When conducted correctly, the main strength of an RCT is its internal validity. In the case of a comparative study, internal validity refers to the confidence with which significant differences emerging between or among groups is due to the intervention, and not confounding variables (Calder, Phillips, & Tybout, 1982). Cook and Campbell (1979) detailed 11 threats to internal validity: history, maturation, testing, instrumentation, statistical regression, selection, mortality, interactions with selection, ambiguity about the direction of causal influence, diffusion or imitation of treatments, compensatory equalization of treatments, compensatory rivalry by respondents receiving less desirable treatments, and resentful demoralization of respondents receiving less desirable treatments. These threats, their meanings, and methods utilized to address these threats in the present research (see Chapter IV) are displayed in Table 16. RCTs may minimize or eliminate these threats to internal validity through the standardized and rigorous nature of their execution, including the use of randomization, manualized interventions, validated and reliable measures, therapist training and supervision, and fidelity monitoring to maximize internal validity (Bellg et al., 2004; Cook & Campbell, 1979). In addition to the threats noted in Table 16, RCTs can also be impacted by type of control condition utilized (Freedland, Mohr, Davidson, & Schwartz, 2011; Mohr et al., 2009). Indeed, in a meta-analyses of RCTs evaluating psychological interventions for depression, outcomes varied significantly based upon the design of the control condition (Mohr et al., 2014). Therefore, control conditions should be considered not only in light of the equity in compensation and resources allotted to each group, but also the type of control condition selected for the execution

of an RCT (Cook & Campbell, 1979; Mohr et al., 2014). When researchers can plausibly eliminate all possible threats to internal validity, confident conclusions about the causal nature of the findings may be made (Cook & Campbell, 1979). This confidence in the findings allows for intervention treatments to be generalized and expanded for use in larger populations.

Limitations. RCTs have limitations to consider both in their interpretation and implementation. First, RCTs are typically resource-intensive and costly (Sanson-Fisher, Bonevski, Green, & D'Este, 2007). Second, due to the standardized nature of their execution to ensure internal validity (see Strengths section above), questions of the external validity of findings have frequently been made (De Los Reyes & Kazdin, 2008; Kazdin, 2008; Pagoto et al., 2007; Rothwell, 2005; Van Spall, Toren, Kiss, & Fowler, 2007; Wilson, 1998). External validity refers to the generalizability of findings to typical clinical practice (Calder et al., 1982). Reasons for these concerns include, but are not limited to: 1) Participants in RCTs meet specific inclusion criteria, making them less likely to have comorbid conditions, changing medication doses, etc. For this reason, RCT participants have often been characterized as having less severe, or complicated presentations than patients in community settings. This claim calls into question the generalizability of RCT findings to community patients with more complex presentations (Hunsley, 2007; Kazdin, 2008). 2) RCT methodology also employs strict guidelines on session length, treatment duration, and goals of treatment (i.e., symptom reduction). As Kazdin (2008) describes, "In clinical practice, much of psychotherapy is not about reaching a destination (eliminating symptoms) as it is about the ride (the process of coping with life)" (p. 147). Clinical practice often allows for more flexibility and adaptability to a patient's changing symptoms and life circumstances, particularly with patient goals frequently changing over the course of treatment (Kazdin, 2008; Sorenson, Gorsuch, & Mintz, 1985). Therefore, employing an

intervention whose efficacy was determined under a strict protocol may be met with hesitation by community practitioners (Pagoto et al., 2007). Concerns about generalizability impact uptake of empirically-supported treatments (i.e., identified as efficacious through an RCT) in community settings. Uptake is also impacted by a 17 year time lag for the conception of an intervention to then be formally evaluated and ultimately translated to community practice (Green, Ottoson, Garcia, & Robert, 2009; Morris, Wooding, & Grant, 2011; Trochim, 2010; Westfall, Mold, & Fagnan, 2007). This length of time is partially due to the structure of clinical research, involving multiple phases (i.e., Phases I, II, III, and IV) and the time required to attain funding between them (Friedman et al., 1998). While attempts to better understand causal factors for this time lag have occurred (Morris et al., 2011), clear causes and solutions have yet to be identified or implemented. The execution and dissemination of RCTs are not without problems and delays. Further, concerns over the generalizability of RCT findings are present and impact dissemination into the community.

Implications for BITs

Increasing Use of Both Methodologies

Stemming from different fields, both usability testing and RCTs are being utilized as means of evaluation for BITs. Usability testing is increasingly being incorporated into the development of mobile (e.g., Ben-Zeev et al., 2013; Kristjánsdóttir et al., 2011; Mansar et al., 2012; Mohr et al., 2015) and web-based BITs (e.g., Voncken-Brewster et al., 2013; Wootten et al., 2014), that are later evaluated via RCTs or other forms of trials. Indeed, mental health professionals interested in the development of apps for mental health are encouraged to engage in multidisciplinary collaboration, and to employ formal usability testing (Price et al., 2014). Further, in a newly proposed design process for optimizing BITs in health systems, usability

testing and RCTs are both included as necessary forms of research (Lyon et al., 2016). Inclusion of both research types in this process enables a multi-faceted approach to adaptability, refinement, and evaluation. The calls for combining the use of both usability testing and RCTs in the evaluation of BITs are becoming more common. Insights into how people use and are impacted by BITs are increased and more nuanced through the execution of both forms of research.

Overcoming Differences

Table 17 outlines differences across the execution and aims of usability testing and RCTs in behavioral medicine. Indeed, typical implementers and consumers of these forms of research have different aims, methodologies, terminology, strengths, and limitations. The differences between usability testing and RCTs do not imply the benefits of one approach over the other. Indeed, the differences exist due to the different goals of these types of research. For example, usability testing typically has relatively small samples, compared to the typically large samples recruited for RCTs. With some controversy (Faulkner, 2003; Sauro, 2010; Woolrych & Cockton, 2001), most usability testing researchers support a sample of five for testing an iteration of a product, provided the evaluation is fairly limited and the user audience is well-defined and represented (Tullis & Albert, 2008). This small sample size tends to identify the majority of usability issues (Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1992). This is in contrast to the sample size powered to detect significant differences between groups in RCTs, often requiring hundreds of participants (Friedman et al., 1998). Despite great differences in sample size, both approaches effectively achieve the respective goals of usability testing and RCTs. Methodological differences that exist in the execution of usability testing and RCTs highlight differing goals and insights achieved through these different forms of research.

Multidisciplinary Communication. Given the differences between usability testing and RCTs, barriers to communication and collaboration are anticipated. Fortunately, the problem of communicating and working across disciplines is not a new issue. Indeed, articles describing multi-disciplinary work and communication spans decades (e.g., Kraut, Egido, & Galegher, 1988). Therefore, insights from past work on multidisciplinary communication may be applied to the present disciplines to enhance future collaboration.

Stowers (2015) recently highlighted common problems and solutions in multidisciplinary work with human factors psychologists and engineers to collaboratively create technological products. A focus on the communication and cooperation of these two disciplines maps well onto anticipated developers and consumers of usability testing and RCTs in BITs. First, in identifying the lack of a unified view of projects among these professionals (de Paula & Barbosa, 2004), possible solutions include: 1) creating scenarios to identify common goals of interaction for a product, and 2) to utilize visual aids when brainstorming (de Paula & Barbosa, 2004; Rosson & Carroll, 2001; Stowers, 2015). Second, to combat differences in terminology, multidisciplinary teams would benefit from explicitly defining terminology when used, or simply stating definitions in practice as opposed to using terminology (Stowers, 2015). This approach would require open dialogue on what "terminology" is and when it is being used, as terminology has been found to be used differently even within the same discipline (Gilb, 2007). Third, when different types of solutions are being sought in research questions (i.e., engineers typically seek concrete solutions, whereas psychologists may seek concrete or soft solutions, based upon the question of human behavior), "if-then" statements could be used to create guidelines. Additionally, the multidisciplinary team can agree to seek solutions in the form of ranges or probabilities in the place of "exact" answers to questions (Stowers, 2015). Finally, engaging in

multidisciplinary communication "too late" into a project, such that projects or products must return to an earlier version to address issues that may have been avoided through collaboration, could be avoided through communication early and throughout development, particularly prior to deployment (Stowers, 2015). Communication and multidisciplinary collaboration between behavioral scientists and engineers is not novel, and can therefore be benefited by previously identified problems and proposed solutions.

Future Directions

Collaboration between the fields of engineering, computer science, and psychology and other behavioral medicine specialties will likely continue to grow in the field of BITs. Indeed, the present studies demonstrate the potential amount of information able to be identified regarding mobile BITs through the use of multiple research methods (see Table 1). Stowers (2015) has already identified some common problems and potential solutions to collaboration and communication among psychologists and engineers. However, there are likely further unidentified barriers and problems in this multidisciplinary work, as well as other solutions likely to be identified and implemented. Future collaborations will therefore benefit from ongoing monitoring and evaluation of the multidisciplinary process.

Summary

A recent call for the refinement and integration of theoretical and research paradigms from multiple disciplines was made (Mohr, Burns, et al., 2013). The aim of this call was to enhance the reach and ability to serve those with mental health needs. Usability testing and RCTs are forms of research typically conducted and consumed by different fields. However, in using both forms of research, the design, development, and deployment of BITs can be improved (Lyon et al., 2016). Ongoing integration of these types of research, and collaboration across

multiple disciplines, will require ongoing refinement. Yet, the efforts may be incredibly fruitful: reaching those needlessly suffering from depression due to barriers to treatment.

Figure 1

Screenshots of Tool and Review Functions of Thought Challenger

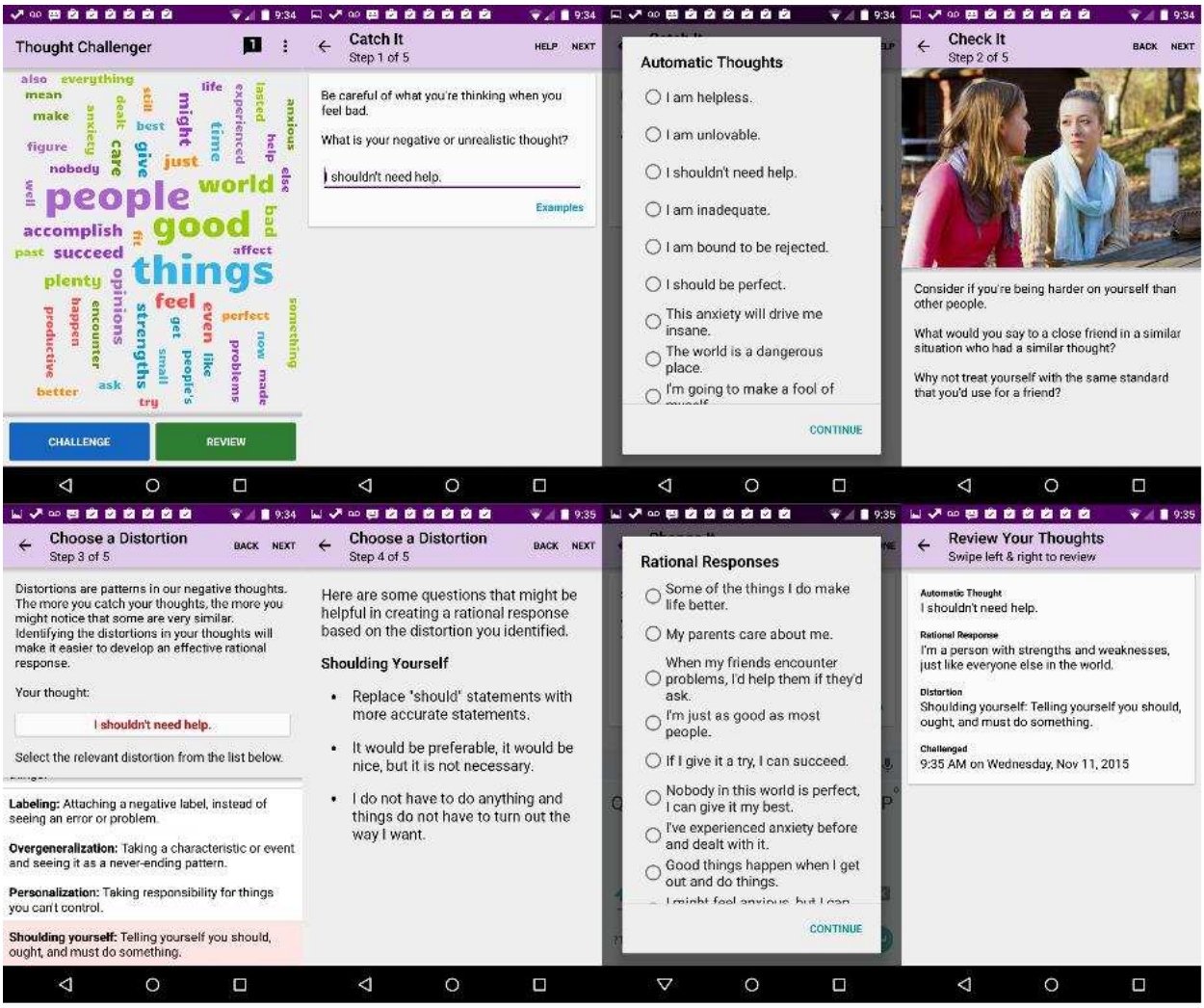


Figure 2

Framework for the Usability Testing Study

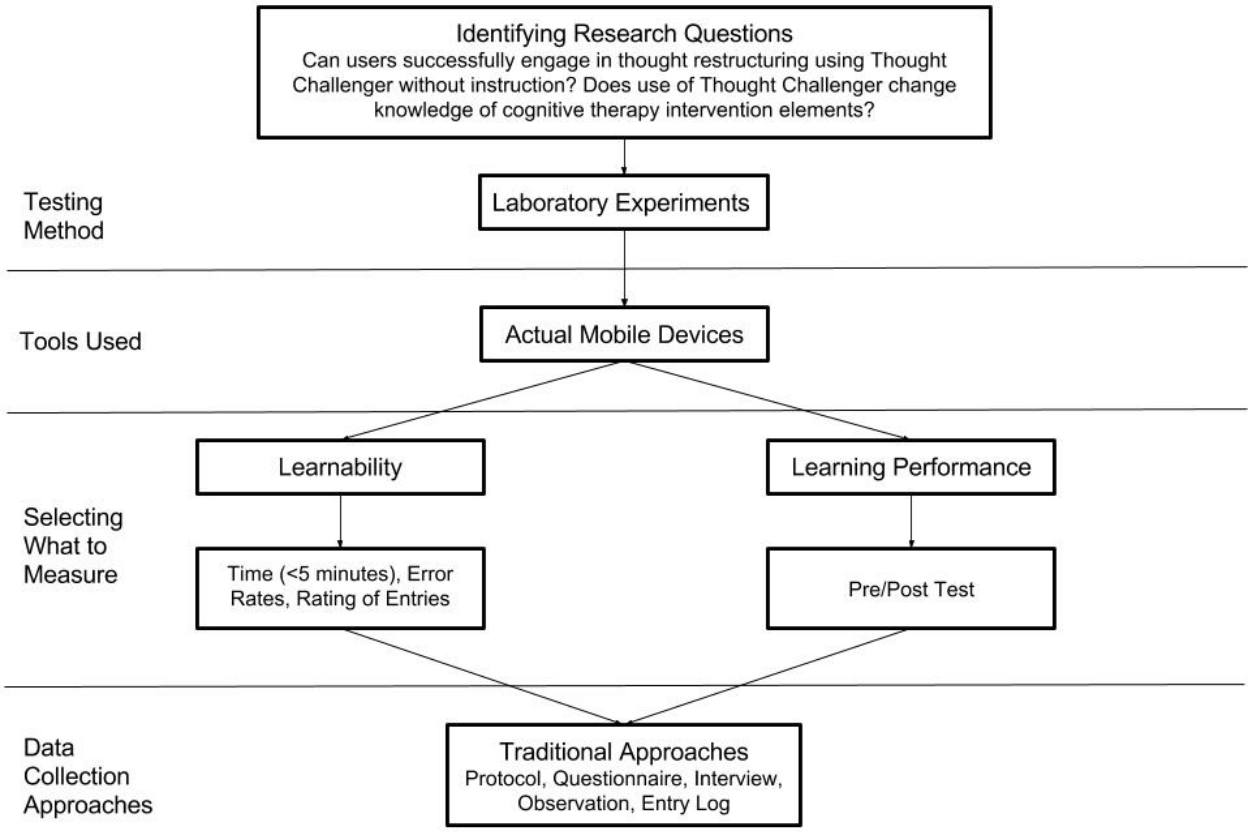


Figure 3

Screenshots of Boost Me Scheduling a New Boost

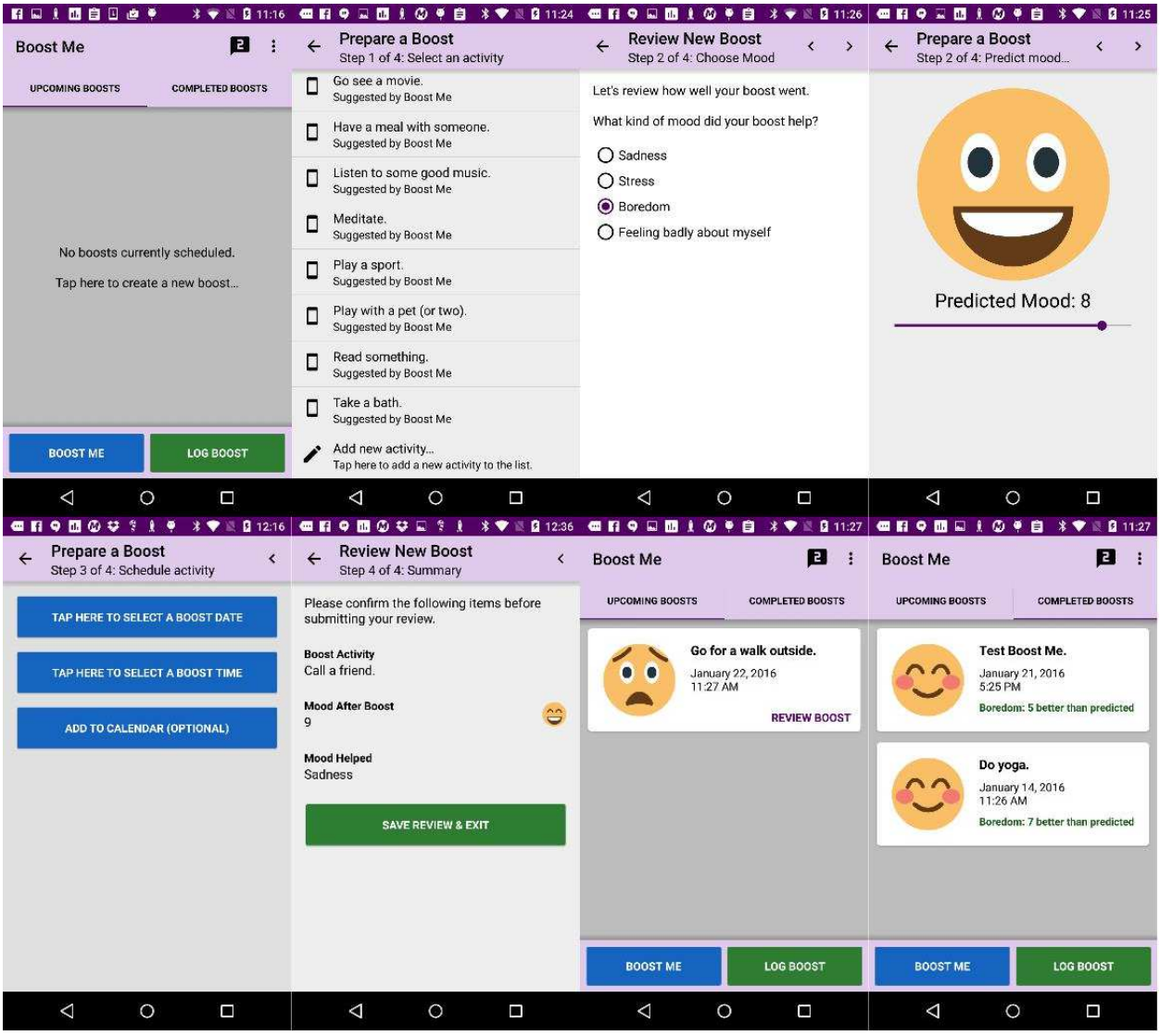
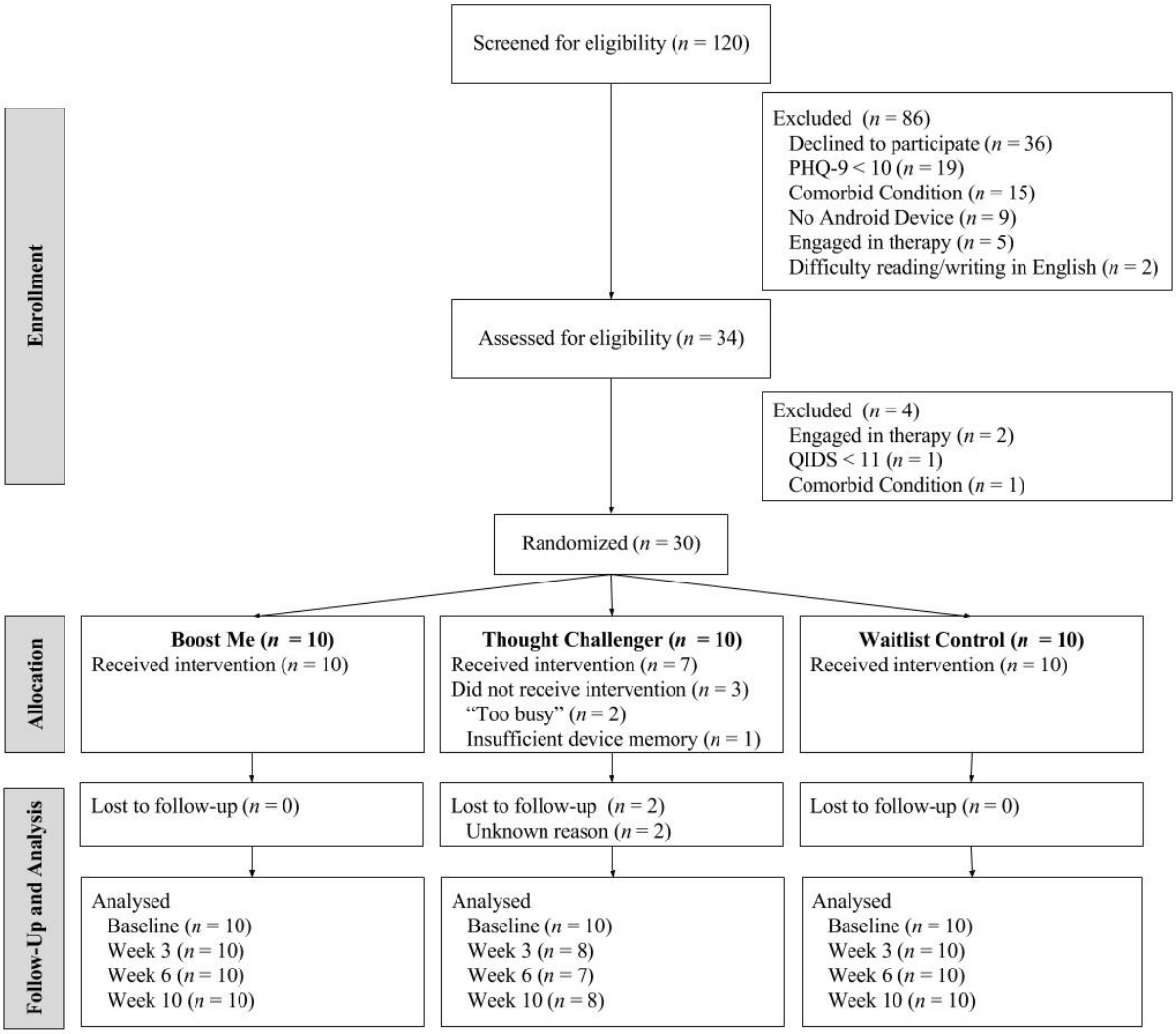


Figure 4

Flow of Participants Through the Trial



Note. PHQ-9 = Patient Health Questionnaire-9; QIDS = Quick Inventory of Depressive Symptomology.

Table 1

Problems, Goals, and Approach for the Present Research

	Problem	Goals	Approach	Study
1	Unknown what barriers users anticipate in using an app for depression	Identify user perceived barriers to the use of apps for depression	Card sorting task to identify barriers with potential end users (n = 20)	Usability Study, Chapter 2
2	Unclear what learning processes and outcomes occur for users of apps for depression	Evaluate the learnability and learning performance of users following initial use of an app for depression	Laboratory usability testing of Thought Challenger evaluating learning in potential end users (n = 20)	Usability Study, Chapter 3
3	Usage and efficacy of treatment apps are varied and poorly defined, individually and comparatively	Evaluate usage and impact on depressive symptoms following use of a BA or CT-informed app	RCT evaluating usage and depressive symptoms in adults with depression (n = 30)	RCT, Chapter 4

Note. BA = Behavioral Activation, CT = Cognitive Therapy, RCT = Randomized Controlled Trial.

Table 2

Card Sorting Sample Characteristics

	PHQ-9 < 10 (*n* = 11)	PHQ-9 ≥ 10 (*n* = 9)	Total (*n* = 20)
Female, *n*(%)	7 (63.6)	8 (88.9)	15 (75)
Age, *M*(SD)	34.5 (10.3)	40.6 (14.0)	37.2 (12.2)
African American, *n*(%)	4 (36.4)	1 (16.7)	5 (25)
Asian, *n*(%)	2 (18.1)	0 (0)	2 (10)
Hispanic Caucasian, *n*(%)	1 (9.1)	0 (0)	1 (5)
Non-Hispanic Caucasian, *n*(%)	5 (45.5)	8 (88.9)	13 (65)
PHQ-9, *M*(SD)	3.8 (3.2)	14.4 (5.8)	8.6 (7.0)
History of Depression, *n*(%)	2 (18.2)	7 (77.8)	9 (45)
History of Anxiety, *n*(%)	2 (18.2)	5 (55.6)	7 (35)

Note. M = mean, SD = standard deviation, PHQ-9 = Patient Health Questionnaire-9.

Table 3

Face-to-Face Delivery Barriers

Group	Variance	Consistency
1	Cost	Cost
2	Lack of insurance coverage	Lack of insurance coverage
	Stigma	Stigma
	Lack of motivation	Lack of motivation
	Concerns about effectiveness	Concerns about effectiveness
	Time for session travel	Time for session travel
	Time for session attendance	Time for session attendance
	Talking about private topics with someone not known	Talking about private topics with someone not known
	Being seen while emotional	Being seen while emotional
3	Discomfort talking about personal issues	Transportation
	Concerns about what friends, family will think	Childcare
	Availability of care	Misfit of therapy to needs
	Not wanting insurance documentation	
4	Distance	Distance
	Want to solve problems on own	Want to solve problems on own
	Time for between session activities	Time for between session activities
	Privacy	Privacy
	Fatigue	Fatigue
	Transportation	Discomfort talking about personal issues
	Misfit of therapy to needs	Availability of care
		Not wanting insurance documentation

Note. Wording in table is identical to the wording the participants viewed on the cards. Groups are listed in order of greatest (1) to smallest (4) barriers. Variance represents clusters formed using mean ranks only (to indicate overall importance); Consistency represents clusters formed using mean ranks and standard deviations (to indicate consistency of importance).

Table 4

App Delivery Barriers

Group	Variance	Consistency
1	Concerns about effectiveness Unsure who has access to data Cost of data package Bugs in the system Wifi access Misfit of features to needs	Concerns about effectiveness Unsure who has access to data Cost of data package Bugs in the system Wifi access Misfit of features to needs
2	Not enough feedback Concerns over lack of guidance	Battery life Concerns over understanding content Time for interaction Notification burden No one caring about how I am doing
3	Lack of human interaction Privacy Lack of motivation Forgetting to use No scheduled time for use Concerns over understanding content No one caring about how I am doing	Lack of human interaction Privacy Lack of motivation Forgetting to use No scheduled time for use Concerns over lack of guidance Not enough feedback
4	Want to solve problems on own Stigma Battery life Time for interaction Notification burden	Want to solve problems on own Stigma

Note. Wording in table is identical to the wording the participants viewed on the cards. Groups are listed in order of greatest (1) to smallest (4) barriers. Variance represents clusters formed using mean ranks only (to indicate overall importance); Consistency represents clusters formed using mean ranks and standard deviations (to indicate consistency of importance).

Table 5

Implications for the Design of Future Apps for Depression Based on User Perceived Barriers

Barrier	Cards	Design Recommendation
Cost	Cost of data package	1. Provide choice of using data package vs. Wifi 2. Explicitly note amount and frequency of data requirements
Privacy and Security	Unsure who has access to data, Privacy	1. Launch clear and concise privacy statement 2. Initiate pop-up request for access to any possible features or data collected from the phone
Efficacy and Functionality	Concerns about effectiveness, Misfit of features to needs, Bugs in the system, Wifi access	1. Provide video testimonials featuring demographically-representative personas 2. Conduct usability testing and quality assurance evaluations prior to deployment 3. Require easily located help button (FAQ and live support connection)
Feedback, guidance, human support	Not enough feedback, Concerns over lack of guidance, Lack of human interaction	1. Provide coach support via phone, text, or messaging 2. Use of algorithms based on context sensing or user behaviors on app

Note. FAQ = Frequently Asked Questions.

Table 6

Dimensions and Attributes of the Revised Bloom's Taxonomy

Knowledge Dimension (Attributes)	Cognitive Process Dimension (Attributes)
A. Factual Knowledge (terminology, specific details/elements)	1. Remember (recognizing, recalling)
B. Conceptual Knowledge (classifications, categories, principles, generalizations, theories, models, structures)	2. Understand (interpreting, exemplifying, classifying, summarizing, inferring, comparing, explaining)
C. Procedural Knowledge (subject-specific skills, algorithms, techniques, and methods; criteria for determining when to use certain procedures)	3. Apply (executing, implementing)
D. Metacognitive Knowledge (cognitive tasks, contextual and conditional elements, self-knowledge, strategic knowledge)	4. Analyze (differentiating, organizing, attributing)
	5. Evaluate (checking, critiquing)
	6. Create (generating, planning, producing)

Note. Adapted from Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*: Allyn & Bacon.

Table 7

Thought Challenger Learning Objectives Mapped onto the Revised Bloom's Taxonomy Table

The Knowledge Dimension	The Cognitive Process Dimension					
	1. Remember	2. Understand	3. Apply	4. Analyze	5. Evaluate	6. Create
A. Factual Knowledge	1	2			5	
B. Conceptual Knowledge		3		5		
C. Procedural Knowledge				4		
D. Metacognitive Knowledge			4			4

Note. Numbers in the body of the table indicate learning objectives for Thought Challenger: 1) Identify specific thoughts that are maladaptive, whether from recent memory or through recognition of similar thought in examples; 2) Understand different types of thought distortions through provided definitions; 3) Classify maladaptive thoughts into distortion categories; 4) Generate specific thoughts that are adaptive, whether from using reflective questions or through identification with an example thought; and 5) Identify common thought patterns through review of entries. Adapted from Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Table 8

Usability Attributes and their Application to Learning Evaluation

	Learnability	Learning Performance
Usability attribute	Level of ease through which a user gains proficiency	Actual impact on performance of a task/acquisition of knowledge
Usability tasks	Complete two attempts at using Thought Challenger tool	Complete a pre and post test of cognitive therapy and skills
Usability measurement	1) Time to complete interactions 2) Error rate 3) Rating of completed thought record	1) Scores on pre and post test
Usability testing learning objectives	1) Identify how user interacts without instruction or didactic material 2) Examine if user learns to use the app within an acceptable time limit, with a low error rate	1) Measure change in knowledge of cognitive therapy skills and concepts following initial use
Application to Revised Bloom's Taxonomy attributes and Thought Challenger learning objectives	-Factual knowledge: Objectives 1, 2, 5 -Conceptual knowledge: Objectives 3, 5 -Remember: Objective 1 -Understand: Objectives 2, 3 -Analyze: Objectives 4, 5 -Create: Objective 4	-Metacognitive knowledge: Objective 4 -Analyze: Objectives 4, 5

Note. Objectives refer to learning objectives for Thought Challenger: 1) Identify specific thoughts that are maladaptive, whether from recent memory or through recognition of similar thought in examples; 2) Understand different types of thought distortions through provided definitions; 3) Classify maladaptive thoughts into distortion categories; 4) Generate specific thoughts that are adaptive, whether from using reflective questions or through identification with an example thought; and 5) Identify common thought patterns through review of entries.

Table 9

Usability Testing Sample Characteristics

	PHQ-9 < 10 (*n* = 11)	PHQ-9 ≥ 10 (*n* = 9)	Total (*n* = 20)
Female, *n*(%)	7 (63.6)	8 (88.9)	15 (75)
Age, *M*(SD)	34.5 (10.3)	40.6 (14.0)	37.2 (12.2)
African American, *n*(%)	4 (36.4)	1 (16.7)	5 (25)
Asian, *n*(%)	2 (18.1)	0 (0)	2 (10)
Hispanic Caucasian, *n*(%)	1 (9.1)	0 (0)	1 (5)
Non-Hispanic Caucasian, *n*(%)	5 (45.5)	8 (88.9)	13 (65)
PHQ-9, *M*(SD)	3.8 (3.2)	14.4 (5.8)	8.6 (7.0)
History of Depression, *n*(%)	2 (18.2)	7 (77.8)	9 (45)
History of Anxiety, *n*(%)	2 (18.2)	5 (55.6)	7 (35)

Note. *M* = mean, SD = standard deviation, PHQ-9 = Patient Health Questionnaire-9.

Table 10

Tool Interaction Completion Times, Median(IQR)

	PHQ-9 < 10	PHQ-9 ≥ 10	Total
Time 1	4:13 (4:01)	3:57 (7:30)	4:05 (4:04)
Time 2	2:08 (1:11)	3:57 (3:40)	2:34 (2:00)

Note. IQR = Interquartile Range; PHQ-9 = Patient Health Questionnaire-9.

Table 11

Cognitive Therapy Pre and Post-Test Scores, Median(IQR)

	PHQ-9 < 10	PHQ-9 ≥ 10	Total
Pre-Test	26.0 (11.0)	29.0 (5.5)	28.5 (11.3)
Post-Test	29.0 (6.0)	32.0 (10.0)	31.0 (6.8)

Note. IQR = Interquartile Range; PHQ-9 = Patient Health Questionnaire-9.

Table 12

Baseline Demographics and Psychiatric Characteristics

Characteristic	Boost Me (*n* =10)	Thought Challenger (*n* = 10)	Waitlist Control (*n* =10)	*p* value
Age, *M*(SD)	35.5 (17.2)	43.1 (11.5)	34.1 (11.0)	.29
Female, *n*(%)	9 (90)	6 (60)	8 (80)	.27
Ethnicity				
Hispanic or Latino, *n*(%)	1 (10)	1 (10)	1 (10)	1.0
Race				
African American, *n*(%)	5 (50)	5 (50)	4 (40)	
White, *n*(%)	5 (50)	4 (40)	6 (60)	.63
Asian, *n*(%)	0 (0)	1 (10)	0 (0)	
Active dose of antidepressant medication, *n*(%)	3 (30)	3 (30)	3 (30)	1.0
PHQ-9, *M*(SD)	15.2 (5.5)	19.2 (5.2)	16.1 (3.8)	.18
QIDS, *M*(SD)	14.6 (2.9)	15.9 (2.2)	14.4 (2.9)	.41
GAD-7, *M*(SD)	10.5 (5.1)	12.6 (6.2)	10.4 (4.4)	.58
Current MDD, *n*(%)	7 (70)	9 (90)	9 (90)	.38
Past MDD, *n*(%)	8 (80)	8 (80)	10 (100)	.32
Passive SI, *n*(%)	6 (60)	9 (90)	9 (90)	.15
Panic Disorder, *n*(%)	3 (30)	2 (20)	1 (10)	.54
Agoraphobia, *n*(%)	3 (30)	6 (60)	5 (50)	.39
Social Phobia Generalized, *n*(%)	2 (20)	0 (0)	1 (10)	.33
Social Phobia Non-Generalized, *n*(%)	2 (20)	0 (0)	0 (0)	.12
Post-Traumatic Stress Disorder, *n*(%)	3 (30)	1 (1)	1 (1)	.38
Alcohol Abuse, *n*(%)	1 (10)	1 (10)	0 (0)	.56
Substance Abuse, *n*(%)	0 (0)	2 (20)	0 (0)	.12
Bulimia Nervosa, *n*(%)	0 (0)	0 (0)	1 (10)	.36
Generalized Anxiety Disorder, *n*(%)	3 (30)	0 (0)	4 (40)	.09

Note. M = Mean; SD = Standard Deviation; PHQ-9 = Patient Health Questionnaire-9; QIDS = Quick Inventory of Depressive Symptomology; GAD-7 = Generalized Anxiety Disorder-7; MDD = Major Depressive Disorder; SI = Suicidal Ideation.

Table 13

Depression Scores Over Time Across Groups, M(SD)

	Baseline	Week 3 (Mid)	Week 6 (EOT)	Week 10 (FU)
Boost Me	15.20 (5.49)	9.60 (4.86)	6.60 (3.95)	8.90 (5.88)
Thought Challenger	17.00 (4.62)	6.14 (3.02)	3.43 (3.82)	5.29 (4.46)
Waitlist Control	16.10 (3.76)	13.60 (5.91)	11.30 (5.58)	11.50 (4.25)

Note. M = Mean; SD = Standard Deviation; EOT = End of treatment; FU = Follow-up.

Table 14

Boost Me and Thought Challenger App Usage

Usage Action, *M* (SD)	Boost Me	Thought Challenger	*p* value
App Launches	97.70 (68.75)	33.50 (37.46)	.02
Week 1	26.60 (26.71)	8.70 (8.68)	.06
Week 2	22.00 (16.15)	8.80 (9.18)	.04
Week 3	14.10 (12.44)	4.40 (6.47)	.04
Week 4	17.50 (14.24)	5.10 (7.71)	.03
Week 5	10.20 (7.70)	3.60 (5.89)	.05
Week 6	7.30 (7.57)	2.90 (5.65)	.16
Events/Thoughts Logged	14.70 (10.07)	8.50 (11.60)	.22
Week 1	1.90 (1.79)	1.40 (1.51)	.51
Week 2	2.90 (1.85)	2.20 (2.35)	.47
Week 3	2.80 (2.62)	1.40 (2.07)	.20
Week 4	4.20 (3.82)	1.20 (2.30)	.05
Week 5	2.20 (2.86)	1.30 (2.26)	.45
Week 6	.70 (1.16)	1.00 (2.11)	.70
Review Launches	7.50 (7.31)	5.40 (4.50)	.45
Week 1	1.50 (1.78)	1.50 (1.43)	1.00
Week 2	2.30 (2.26)	2.20 (2.30)	.92
Week 3	1.10 (1.66)	.60 (.84)	.41
Week 4	.90 (1.91)	.70 (1.57)	.76
Week 5	.80 (1.87)	.40 (.70)	.54
Week 6	.90 (1.91)	0 (0)	.15
Completed Scheduled Activity	8.20 (10.05)	-	-
Week 1	.90 (1.29)	-	-
Week 2	1.80 (1.87)	-	-
Week 3	1.60 (2.56)	-	-
Week 4	2.00 (3.46)	-	-
Week 5	1.30 (2.75)	-	-
Week 6	.60 (1.27)	-	-

Note. M = Mean; SD = Standard Deviation.

Table 15

Problems, Goals, and Findings for the Present Research

	Problems	Goals	Findings
1	Unknown what barriers users anticipate in using an app for depression	Identify user perceived barriers to the use of apps for depression	Top app barriers: concerns over intervention efficacy, app functioning, privacy, cost, and lack of guidance and tailored feedback
2	Unclear what learning processes and outcomes occur for users of apps for depression	Evaluate the learnability and learning performance of users following initial use of an app for depression	Thought Challenger: learnable at acceptable time, low error rate, promotes effective execution of thought restructuring, CT knowledge/skills improve significantly
3	Usage and efficacy of treatment apps are varied and poorly defined, individually and comparatively	Evaluate usage and impact on depressive symptoms following use of a BA or CT-informed app	Boost Me used significantly more than Thought Challenger; significant differences in depression scores over time between Thought Challenger, Waitlist control participants

Note. BA = Behavioral Activation, CT = Cognitive Therapy.

Table 16

Threats to Internal Validity and How They Were Accounted for in the Current Research

Threat	Meaning	Method Employed to Account for Threat
History	Observed effect due to an event that is not the intervention takes place between the pre and post test	Randomization, Blocked randomization design
Maturation	Observed effect due to participants aging, gaining more experience between the pre and post test	Randomization
Testing	Familiarity with a test increases performance over time	Use of symptom-based measure
Instrumentation	Observed effect due to a change in the assessment instrument	Randomization, Use of consistent measures over time and across groups
Statistical Regression	Observed effect due to participants being classified into groups on the basis of pretest scores	Randomization
Selection	Observed effect due to differences between participants assigned to one group, compared to the other	Randomization
Mortality	Observed effect due to the types of participants who drop out of treatment	Randomization
Interactions with Selection	Observed effect due to interactions of other threats with the selection of items (i.e., selection-maturation, selection-history, selection-instrumentation)	Randomization
Ambiguity about the Direction of Causal Influence	Unclear if observed effect is A causing B, or B causing A	Use of a clear order of temporal precedence
Diffusion, Imitation of Treatments	Observed effect due to participants from different groups sharing their experiences with one another	National recruitment, Maintenance of confidentiality
Compensatory Equalization of Treatments	Observed effect due to inequity of goods, services among groups	Equal subject payment, Eventual access to interventions for all groups
Compensatory Rivalry by Respondents Receiving Less Desirable Treatments	Observed effect due to the control group feeling motivated to reduce, reverse the expected difference	Blinded to hypotheses

Resentful Demoralization of Respondents Receiving Less Desirable Treatments	Observed effect due to the control group feeling resentment due to compensatory rivalry	Equal subject payment, Eventual access to interventions for all groups

Table 17

Overview of Usability Testing and RCTs

	Usability Testing	RCTs
Definition	Systematic observation of a planned task or scenario carried out by an actual or potential user	Prospective experiment testing the efficacy of a treatment against a control condition
Importance	Ensure that a technology is intuitive and easy to use, inform decisions; can impact people on a scale ranging from convenience to life saving	Ensure that the effect of an intervention is significantly better than no treatment, or a previously established treatment; can generalize to interventions for larger populations
Outcome Data	Measureable, quantifiable metrics revealing information about interaction of a user and a technology	Valid, reliable comparable quantities revealing information about a treatment's efficacy
Typical Consumers	Engineering, Computer Science via Conference Papers	Behavioral Sciences, Public Health via Journal Articles
Strengths	Prevent costly errors, demonstrate impact of design changes, quick, specificity, increased likelihood of use	If conducted correctly, high internal validity
Limitations	Possible biases resulting from: participants, tasks, methods, artifacts, environment, and moderator	Resource intensive, questions of external validity, time lag in translation

Note. RCTs = randomized controlled trials.

References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology, 68*(2), 331-339.

Aderka, I. M., Nickerson, A., Boe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology, 80*(1), 93-101. doi: 10.1037/a0026455

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-V*. Arlington, VA: American Psyhicatric Publishing.

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York,: Academic Press.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*: Allyn & Bacon.

Andersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: A meta-analysis. *Cognitive Behaviour Therapy, 38*(4), 196-205. doi: 10.1080/16506070903318960

Andrews, G., Cuijpers, P., Craske, M. G., McEvoy, P., & Titov, N. (2010). Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: A meta-analysis. *PLoS One, 5*(10), e13196. doi: 10.1371/journal.pone.0013196

Appiah, O. (2006). Rich media, poor media: The impact of audio/video vs. text/picture testimonial ads on browsers' evaluations of commercial web sites and online products. *Journal of Current Issues & Research in Advertising, 28*(1), 73-86.

Årsand, E., Frøisland, D. H., Skrøvseth, S. O., Chomutare, T., Tatara, N., Hartvigsen, G., & Tufano, J. T. (2012). Mobile health applications to assist patients with diabetes: Lessons

learned and design implications. *Journal of Diabetes Science and Technology, 6*(5), 1197-1206.

Backenstrass, M., Frank, A., Joest, K., Hingmann, S., Mundt, C., & Kronmuller, K. T. (2005). A comparative study of nonspecific depressive symptoms and minor depression regarding functional impairment and associated characteristics in primary care. *Comprehensive Psychiatry, 47*(1), 35-41.

Barak, A., Klein, B., & Proudfoot, J. G. (2009). Defining internet-supported therapeutic interventions. *Annals of Behavioral Medicine, 38*(1), 4-17. doi: 10.1007/s12160-009-9130-7

Barber, J. P., & DeRubeis, R. J. (1989). On second thought: Where the action is in cognitive therapy for depression. *Cognitive Therapy and Research, 13*(5), 441-457.

Beatty, L., & Binnion, C. (2016). A systematic review of predictors of, and reasons for, adherence to online psychological interventions. *International Journal of Behavioral Medicine*. doi: 10.1007/s12529-016-9556-9

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Beck, J. S. (1995). *Cognitive therapy*. New York: Guilford Press.

Beck, J. S. (2011). *Cognitive behavior therapy* (Second Edition ed.). New York: Guilford Press.

Bedford, A. (2014). Instructional overlays and coach marks for mobile apps: Nielsen Norman Group: Evidence-Based User Experience Research, Training, and Consulting.

Bedi, N., Chilvers, C., Churchill, R., Dewey, M., Duggan, C., Fielding, K., . . . Williams, I. (2000). Assessing effectiveness of treatment of depression in primary care. Partially randomised preference trial. *British Journal of Psychiatry, 177*, 312-318.

Behnken, A., Schoning, S., Gerss, J., Konrad, C., de Jong-Meyer, R., Zwanzger, P., & Arolt, V. (2010). Persistent non-verbal memory impairment in remitted major depression - caused by encoding deficits? *Journal of Affective Disorders, 122*(1-2), 144-148. doi: 10.1016/j.jad.2009.07.010

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., . . . Treatment Fidelity Workgroup of the N. I. H. Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health psychology, 23*(5), 443-451. doi: 10.1037/0278-6133.23.5.443

Ben-Zeev, D., Kaiser, S. M., Brenner, C. J., Begale, M., Duffecy, J., & Mohr, D. C. (2013). Development and usability testing of FOCUS: A smartphone system for self-management of schizophrenia. *Psychiatric Rehabilitation Journal, 36*(4), 289-296. doi: 10.1037/prj0000019

Berger, T., Hammerli, K., Gubser, N., Andersson, G., & Caspar, F. (2011). Internet-based treatment of depression: A randomized controlled trial comparing guided with unguided self-help. *Cognitive Behaviour Therapy, 40*(4), 251-266. doi: 10.1080/16506073.2011.616531

Bevan, N. (2009). Usability *Encyclopedia of Database Systems* (pp. 3247-3251): Springer.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi: 10.1146/annurev-psych-113011-143823

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives*. New York: David McKay Company.

Brody, D. S., Khaliq, A. A., & Thompson, T. L., 2nd. (1997). Patients' perspectives on the management of emotional distress in primary care settings. *Journal of General Internal Medicine, 12*(7), 403-406.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas & B. A. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189-194). London, UK: Taylor & Francis.

Buijink, A. W., Visser, B. J., & Marshall, L. (2013). Medical apps for smartphones: Lack of evidence undermines quality and safety. *Evid Based Med, 18*(3), 90-92. doi: 10.1136/eb-2012-100885

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research, 13*(3), e55. doi: 10.2196/jmir.1838

Burns, M. N., & Mohr, D. C. (2013). eHealth and behavioral intervention technologies. In M. D. Gellman & J. R. Turner (Eds.), *Encyclopedia of behavioral medicine* (pp. 659-664). New York: Springer.

Busch, A. M., Kanter, J. W., Landes, S. J., & Kohlenberg, R. J. (2006). Sudden gains and outcome: A broader temporal analysis of cognitive therapy for depression. *Behavior Therapy, 37*(1), 61-68. doi: 10.1016/j.beth.2005.04.002

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31. doi: 10.1016/j.cpr.2005.07.003

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, 240-244.

Campbell, S., & Macqueen, G. (2004). The role of the hippocampus in the pathophysiology of major depression. *Journal of Psychiatry & Neuroscience, 29*(6), 417-426.

Carroll, J. M., & Rosson, M. B. (1987). The paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction*. Cambridge, MA: MIT Press.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52*, 685-716. doi: 10.1146/annurev.psych.52.1.685

Chiu, T. M., & Eysenbach, G. (2010). Stages of use: Consideration, initiation, utilization, and outcomes of an internet-mediated intervention. *BMC Medicineical Informatics and Decision Making, 10*(73). doi: 10.1186/1472-6947-10-73

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research, 11*(2), e13. doi: 10.2196/jmir.1194

Christensen, H., Griffiths, K. M., & Korten, A. (2002). Web-based cognitive behavior therapy: analysis of site usage and changes in depression and anxiety scores. *Journal of Medical Internet Research, 4*(1), e3. doi: 10.2196/jmir.4.1.e3

Christensen, H., Griffiths, K. M., Korten, A. E., Brittliffe, K., & Groves, C. (2004). A comparison of changes in anxiety and depression symptoms of spontaneous users and trial participants of a cognitive behavior therapy website. *Journal of Medical Internet Research, 6*(4), e46. doi: 10.2196/jmir.6.4.e46

Christensen, H., & Hickie, I. B. (2010). E-mental health: A new era in delivery of mental health services. *Medical Journal of Australia, 192*(11), S2.

Churchill, R., Khaira, M., Gretton, V., Chilvers, C., Dewey, M., Duggan, C., . . . Group, Antidepressants in Primary Care Study. (2000). Treating depression in general practice: Factors affecting patients' treatment preferences. *British Journal of General Practice, 50*(460), 905-906.

Clarke, G., & Yarborough, B. J. (2013). Evaluating the promise of health IT to enhance/expand the reach of mental health services. *General Hospital Psychiatry, 35*(4), 339-344. doi: 10.1016/j.genhosppsych.2013.03.013

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cooper, L. A., Gonzales, J. J., Gallo, J. J., Rost, K. M., Meredith, L. S., Rubenstein, L. V., . . . Ford, D. E. (2003). The acceptability of treatment for depression among African-American, Hispanic, and white primary care patients. *Medical Care, 41*(4), 479-489.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry, 58*(7), 376-385.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology, 76*(6), 909-922. doi: 10.1037/a0013075

Cuijpers, P., van Straten, A., & Warmerdam, L. (2007). Behavioral activation treatments of depression: A meta-analysis. *Clinical Psychology Review, 27*(3), 318-326.

Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., & Penninx, B. W. (2013). Differential mortality rates in major and subthreshold depression: Meta-analysis of studies that measured both. *The British Journal of Psychiatry, 202*(1), 22-27. doi: 10.1192/bjp.bp.112.112169

Darkins, A., Ryan, P., Kobb, R., Foster, L., Edmonson, E., Wakefield, B., & Lancaster, A. E. (2008). Care coordination/home telehealth: The systematic implementation of health informatics, home telehealth, and disease management to support the care of veteran patients with chronic conditions. *Telemedicine and e-Health, 14*(10), 1118-1126.

De Los Reyes, A., & Kazdin, A. E. (2008). When the evidence says, "Yes, No, and Maybe So": Attending to and Interpreting Inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science, 17*(1), 47-51. doi: 10.1111/j.1467-8721.2008.00546.x

de Paula, M., & Barbosa, S. D. J. (2004). *Using an interaction model to support communication among HCI design team members from multidisciplinary backgrounds*. Paper presented at the Workshop on Human Factors in Computer Systems.

Dehling, T., Gao, F., Schneider, S., & Sunyaev, A. (2015). Exploring the far side of mobile health: Information security and privacy of mobile health apps on iOS and Android. *JMIR mHealth uHealth, 3*(1), e8. doi: 10.2196/mhealth.3672

Derbyshire, E., & Dancey, D. (2013). Smartphone medical applications for women's health: What is the evidence-base and feedback? *International Journal of Telemedicine and Applications*. doi: 10.1155/2013/782074

Desoete, A. (2007). Evaluating and improving the mathematics teaching-learning process through metacognition. *Electronic Journal of Research in Educational Psychology, 5*(3), 705-730.

DiMatteo, M. R., Lepper, H. S., & Croghan, T. W. (2000). Depression is a risk factor for noncompliance with medical treatment: Meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of Internal Medicine, 160*(14), 2101-2107.

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., . . . Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology, 74*(4), 658-670. doi: 10.1037/0022-006X.74.4.658

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57*(3), 414-419.

Dobson, K. S., Hollon, S. D., Dimidjian, S., Schmaling, K. B., Kohlenberg, R. J., Gallop, R. J., . . . Jacobson, N. S. (2008). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the prevention of relapse and recurrence in major depression. *Journal of Consulting and Clinical Psychology, 76*(3), 468-477. doi: 10.1037/0022-006X.76.3.468

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M. R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: A systematic review. *Journal of Medical Internet Research, 15*(11), e247. doi: 10.2196/jmir.2791

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., & Glozier, N. (2011). A systematic review of the impact of adherence on the effectiveness of e-therapies. *Journal of Medical Internet Research, 13*(3), e52. doi: 10.2196/jmir.1772

Dwight-Johnson, M., Sherbourne, C. D., Liao, D., & Wells, K. B. (2000). Treatment preferences among depressed primary care patients. *Journal of General Internal Medicine, 15*(8), 527-534.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods Instruments & Computers, 35*(3), 379-383. doi: Doi 10.3758/Bf03195514

Fava, M., Rankin, M. A., Wright, E. C., Alpert, J. E., Nierenberg, A. A., Pava, J., & Rosenbaum, J. F. (2000). Anxiety disorders in major depression. *Comprehensive Psychiatry, 41*(2), 97-102.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., . . . Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010. *PLoS Med, 10*(11), e1001547. doi: 10.1371/journal.pmed.1001547

Ferrari, A. J., Somerville, A. J., Baxter, A. J., Norman, R., Patten, S. B., Vos, T., & Whiteford, H. A. (2013). Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature. *Psychological Medicine, 43*(3), 471-481. doi: 10.1017/S0033291712001511

Flood, D., Harrison, R., Iacob, C., & Duce, D. (2012). Evaluating mobile applications: A spreadsheet case study. *International Journal of Mobile Human Computer Interaction, 4*(4), 37-65.

Freedland, K. E., Mohr, D. C., Davidson, K. W., & Schwartz, J. E. (2011). Usual and unusual care: Existing practice control groups in randomized controlled trials of behavioral interventions. *Psychosom Med, 73*(4), 323-335. doi: 10.1097/PSY.0b013e318218e1fb

Friedman, L. M., Furberg, C., & DeMets, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer.

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013). *Why people hate your app: Making sense of user feedback in a mobile app store.* Paper presented at the 19th Annual Conference Meeting of the International Conference on Knowledge Discovery and Data Mining.

Garner, R., & Alexander, P. A. (1989). Metacognition: Answered and unanswered questions. *Educational Psychologist, 24*(2), 143-158.

Genz, A., Kirk, G., Piggott, D., Mehta, S. H., Linas, B. S., & Westergaard, R. P. (2015). Uptake and acceptability of information and communication technology in a community-based cohort of people who inject drugs: Implications for mobile health interventions. *JMIR mHealth uHealth, 3*(2), e70. doi: 10.2196/mhealth.3437

Gilb, T. (2007). *Undergraduate basics for Systems Engineering (SE), using the principles, measures, concepts and processes of planguage.* Paper presented at the INCOSE International Symposium.

Gilbody, S., Littlewood, E., Hewitt, C., Brierley, G., Tharmanathan, P., Araya, R., . . . Team, Reeact. (2015). Computerised cognitive behaviour therapy (cCBT) as treatment for

depression in primary care (REEACT trial): Large scale pragmatic randomised controlled trial. *BMJ, 351*, h5627. doi: 10.1136/bmj.h5627

Gilliam, M. L., Martins, S. L., Bartlett, E., Mistretta, S. Q., & Holl, J. L. (2014). Development and testing of an iOS waiting room "app" for contraceptive counseling in a Title X family planning clinic. *American Journal of Obstetrics and Gynecology, 211*(5), 481. e481-481. e488.

Glick, G., Druss, B., Pina, J., Lally, C., & Conde, M. (2015). Use of mobile technology in a community mental health setting. *Journal of Telemedicine and Telecare*, 1357633X15613236.

Gonzalez, H. M., Vega, W. A., Williams, D. R., Tarraf, W., West, B. T., & Neighbors, H. W. (2010). Depression care in the United States: Too little for too few. *Archives of General Psychiatry, 67*(1), 37-46. doi: 10.1001/archgenpsychiatry.2009.168

Green, L. W., Ottoson, J., Garcia, C., & Robert, H. (2009). Diffusion theory and knowledge dissemination, utilization, and integration in public health. *Annu Rev Public Health, 30*, 151.

Gumport, N. B., Williams, J. J., & Harvey, A. G. (2015). Learning cognitive behavior therapy. *Journal of behavior therapy and experimental psychiatry, 48*, 164-169. doi: 10.1016/j.jbtep.2015.03.015

Hardy, G. E., Cahill, J., Stiles, W. B., Ispan, C., Macaskill, N., & Barkham, M. (2005). Sudden gains in cognitive therapy for depression: A replication and extension. *Journal of Consulting and Clinical Psychology, 73*(1), 59-67. doi: 10.1037/0022-006X.73.1.59

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow

process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377-381.

Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: Literature review and rationale for a new usability model. *Journal of Interaction Science, 1*(1), 1-16.

Harvey, A. G., Lee, J., Williams, J., Hollon, S. D., Walker, M. P., Thompson, M. A., & Smith, R. (2014). Improving outcome of psychosocial treatments by enhancing memory and learning. *Perspectives on Psychological Science, 9*(2), 161-179. doi: 10.1177/1745691614521781

He, D., Naveed, M., Gunter, C. A., & Nahrstedt, K. (2014). Security concerns in Android mHealth apps. *AMIA Annual Symposium Proceedings*, 645-654.

Helander, E., Kaipainen, K., Korhonen, I., & Wansink, B. (2014). Factors related to sustained use of a free mobile app for dietary self-monitoring with photography and peer feedback: Retrospective cohort study. *Journal of Medical Internet Research, 16*(4), e109. doi: 10.2196/jmir.3084

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the Patient Health Questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine, 21*(6), 547-552. doi: 10.1111/j.1525-1497.2006.00409.x

Huckvale, K., Car, M., Morrison, C., & Car, J. (2012). Apps for asthma self-management: A systematic assessment of content and tools. *BMC Medicine, 10*, 144. doi: 10.1186/1741-7015-10-144

Hundt, N. E., Mignogna, J., Underhill, C., & Cully, J. A. (2013). The relationship between use of CBT skills and depression treatment outcome: A theoretical and methodological review of the literature. *Behavior Therapy, 44*(1), 12-26.

Hunnicutt-Ferguson, K., Hoxha, D., & Gollan, J. (2012). Exploring sudden gains in behavioral activation therapy for major depressive disorder. *Behaviour Research Therapy, 50*(3), 223-230. doi: http://dx.doi.org/10.1016/j.brat.2012.01.005

Hunsley, J. (2007). Addressing key challenges in evidence-based practice in psychology. *Professional Psychology: Research and Practice, 38*(2), 113.

IMS Institute for Healthcare Informatics. (2015). Patient adoption of mHealth: Use, evidence and remaining barriers to mainstream acceptance. Parisppany, NJ: IMS Institute for Healthcare Informatics.

James, D. C., & Harville, C. (2015). Barriers and motivators to participating in mHealth research among African American men. *American Journal of Men's Health*. doi: 10.1177/1557988315620276

Jarrett, R. B., Vittengl, J. R., Clark, L. A., & Thase, M. E. (2011). Skills of cognitive therapy (SoCT): A new measure of patients' comprehension and use. *Psychological Assessment, 23*(3), 578-586. doi: 10.1037/a0022485

Jones, M., Ebert, D. D., Jacobi, C., Beintner, I., Berger, T., Gorlich, D., . . . Botella, C. (2015). Why didn't patients use it? Engagement is the real story in Gilbody et al. (2015), not effectiveness. *BMJ, 351*, h5627.

Judd, L. L., Paulus, M. P., Wells, K. B., & Rapaport, M. H. (1996). Socioeconomic burden of subsyndromal depressive symptoms and major depression in a sample of the general population. *The American Journal of Psychiatry, 153*(11), 1411-1417.

Judd, L. L., Paulus, M. P., & Zeller, P. (1999). The role of residual subthreshold depressive symptoms in early episode relapse in unipolar major depressive disorder. *Archives of General Psychiatry, 56*(8), 764-765.

Karekla, M., Lundgren, J. D., & Forsyth, J. P. (2004). A survey of graduate training in empirically supported and manualized treatments: A preliminary report. *Cognitive and Behavioral Practice, 11*(2), 230-242.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*(3), 146-159. doi: 10.1037/0003-066X.63.3.146

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science, 6*(1), 21-37.

Kessler, R. C., Zhao, S., Blazer, D. G., & Swartz, M. (1997). Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *Journal of Affective Disorders, 45*(1-2), 19-30.

Khalid, H., Shihab, E., Nagappan, M., & Hassan, A. (2014). What do mobile app users complain about? A study on free iOS apps. *IEEE Software, 32*(3), 70-77. doi: 10.1109/MS.2014.50

Kiili, K. (2002). *Evaluating WAP usability:" What usability?".* Paper presented at the Wireless and Mobile Technologies in Education.

Kim, B., Park, H., & Baek, Y. (2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education, 52*(4), 800-810.

Koole, M. L. (2009). A model for framing mobile learning. In M. Ally (Ed.), *Mobile learning: Transforming the delivery of education and training*. Edmonton, AB: AU Press.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Kraut, R., Egido, C., & Galegher, J. (1988). *Patterns of contact and communication in scientific research collaboration.* Paper presented at the Proceedings of the 1988 ACM conference on Computer-supported cooperative work.

Krebs, P., & Duncan, D. T. (2015). Health app use among US mobile phone owners: A national survey. *JMIR mHealth uHealth, 3*(4), e401.

Kristjánsdóttir, Ó. B., Fors, E. A., Eide, E., Finset, A., van Dulmen, S., Wigers, S. H., & Eide, H. (2011). Written online situational feedback via mobile phone to support self-management of chronic widespread pain: A usability study of a web-based intervention. *BMC Musculoskeletal Disorders, 12*(1), 1-9. doi: 10.1186/1471-2474-12-51

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 1-7.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613.

Laborde-Lahoz, P., El-Gabalawy, R., Kinley, J., Kirwin, P. D., Sareen, J., & Pietrzak, R. H. (2015). Subsyndromal depression among older adults in the USA: Prevalence, comorbidity, and risk for new-onset psychiatric disorders in late life. *International Journal of Geriatric Psychiatry, 30*(7), 677-685. doi: 10.1002/gps.4204

Lattie, E. G., Schueller, S. M., Sargent, E., Stiles-Shields, C., Tomasino, K. N., Corden, M. E., . . . Mohr, D. C. (In Press). Uptake and usage of Intellicare: A publicly available suite of mental health and well-being apps. *Internet Interventions.*

Lee, R. S., Hermens, D. F., Porter, M. A., & Redoblado-Hodge, M. A. (2012). A meta-analysis of cognitive deficits in first-episode major depressive disorder. *Journal of Affective Disorders, 140*(2), 113-124. doi: 10.1016/j.jad.2011.10.023

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36*(2), 368-378.

Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012). *Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing.* Paper presented at the 2012 ACM Conference on Ubiquitous Computing.

Lin, J., Liu, B., Sadeh, N., & Hong, J. I. (2014). *Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings.* Paper presented at the Symposium On Usable Privacy and Security.

Lowe, B., Kroenke, K., Herzog, W., & Grafe, K. (2004). Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the Patient Health Questionnaire (PHQ-9). *Journal of Affective Disorders, 81*(1), 61-66. doi: 10.1016/S0165-0327(03)00198-8

Luchini, K., Quintana, C., & Soloway, E. (2003). *Pocket PiCoMap: A case study in designing and assessing a handheld concept mapping tool for learners.* Paper presented at the SIGCHI Conference on Human Factors in Computing Systems.

Lyon, A. R., Wasse, J. K., Ludwig, K., Zachry, M., Bruns, E. J., Unutzer, J., & McCauley, E. (2016). The Contextualized Technology Adaptation Process (CTAP): Optimizing health information technology to Improve mental health systems. *Adm Policy Ment Health, 43*(3), 394-409. doi: 10.1007/s10488-015-0637-x

MacQueen, G. M., Campbell, S., McEwen, B. S., Macdonald, K., Amano, S., Joffe, R. T., . . . Young, L. T. (2003). Course of illness, hippocampal function, and hippocampal volume in major depression. *Proceedings of the National Academy of Sciences, 100*(3), 1387-1392. doi: 10.1073/pnas.0337481100

Maher, C. A., Lewis, L. K., Ferrar, K., Marshall, S., De Bourdeaudhuij, I., & Vandelanotte, C. (2014). Are health behavior change interventions that use online social networks effective? A systematic review. *Journal of Medical Internet Research, 16*(2), e40. doi: 10.2196/jmir.2952

Mansar, S. L., Jariwala, S., Shahzad, M., Anggraini, A., Behih, N., & AlZeyara, A. (2012). A usability testing experiment for a localized weight loss mobile application. *Procedia Technology, 5*, 839-848. doi: http://dx.doi.org/10.1016/j.protcy.2012.09.093

Martell, C. R., Dimidjian, S., & Herman-Dunn, R. (2010). *Behavioral activation for depression: A clinician's guide*. New York: The Guilford Press.

Martinez-Perez, B., de la Torre-Diez, I., & Lopez-Coronado, M. (2013). Mobile health applications for the most prevalent conditions by the World Health Organization: Review and analysis. *Journal of Medical Internet Research, 15*(6), e120. doi: 10.2196/jmir.2600

Martinez-Perez, B., de la Torre-Diez, I., & Lopez-Coronado, M. (2015). Privacy and security in mobile health apps: A review and recommendations. *Journal of Medical Systems, 39*(1), 181. doi: 10.1007/s10916-014-0181-3

Melville, K. M., Casey, L. M., & Kavanagh, D. J. (2010). Dropout from Internet-based treatment for psychological disorders. *British Journal of Clinical Psychology, 49*(Pt 4), 455-471. doi: 10.1348/014466509X472138

Menke, R., & Flynn, H. (2009). Relationships between stigma, depression, and treatment in white and African American primary care patients. *The Journal of Nervous and Mental Disease, 197*(6), 407-411.

Merriam-Webster.com.). learning. Retrieved January 28, 2016, from http://www.merriam-webster.com/dictionary/learning

Miner, A., Schueller, S. M., Lattie, E. G., & Mohr, D. C. (2015). Creation and validation of the cognitive behavioral therapy skills assessment in a depression trial. *Psychiatry Research, 230*(3), 819-825.

Mohr, D. C., Burns, M. N., Schueller, S. M., Clarke, G., & Klinkman, M. (2013). Behavioral intervention technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry, 35*(4), 332-338. doi: 10.1016/j.genhosppsych.2013.03.008

Mohr, D. C., Cheung, K., Schueller, S. M., Hendricks Brown, C., & Duan, N. (2013). Continuous evaluation of evolving behavioral intervention technologies. *American Journal of Preventive Medicine, 45*(4), 517-523. doi: 10.1016/j.amepre.2013.06.006

Mohr, D. C., Cuijpers, P., & Lehman, K. (2011). Supportive accountability: A model for providing human support to enhance adherence to eHealth interventions. *Journal of Medical Internet Research, 13*(1), e30.

Mohr, D. C., Duffecy, J., Ho, J., Kwasny, M., Cai, X., Burns, M. N., & Begale, M. (2013). A randomized controlled trial evaluating a manualized TeleCoaching protocol for improving adherence to a web-based intervention for the treatment of depression. *PLoS One, 8*(8), e70086. doi: 10.1371/journal.pone.0070086

Mohr, D. C., Hart, S. L., Howard, I., Julian, L., Vella, L., Catledge, C., & Feldman, M. D. (2006). Barriers to psychotherapy among depressed and nondepressed primary care patients. *Annals of Behavioral Medicine, 32*(3), 254-258. doi: 10.1207/s15324796abm3203_12

Mohr, D. C., Ho, J., Duffecy, J., Baron, K. G., Lehman, K. A., Jin, L., & Reifler, D. (2010). Perceived barriers to psychological treatments and their relationship to depression. *Journal of Clinical Psychology, 66*(4), 394-409. doi: 10.1002/jclp.20659

Mohr, D. C., Ho, J., Hart, T. L., Baron, K. G., Berendsen, M., Beckner, V., . . . Duffecy, J. (2014). Control condition design and implementation features in controlled trials: A meta-analysis of trials evaluating psychotherapy for depression. *Translational Behavioral Medicine, 4*(4), 407-423. doi: 10.1007/s13142-014-0262-3

Mohr, D. C., Montague, E., Stiles-Shields, C., Kaiser, S. M., Brenner, C., Carty-Fickes, E., . . . Duffecy, J. (2015). Medlink: A mobile intervention to address failure points in the treatment of depression in general medicine. *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 100-107.

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

Möller-Leimkühler, A. M. (2002). Barriers to help-seeking by men: A review of sociocultural and clinical literature with particular reference to depression. *Journal of Affective Disorders, 71*(1), 1-9.

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*(12), 510-520.

Mueller, T. I., Leon, A. C., Keller, M. B., Solomon, D. A., Endicott, J., Coryell, W., . . . Maser, J. D. (1999). Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *The American Journal of Psychiatry, 156*(7), 1000-1006. doi: 10.1176/ajp.156.7.1000

Murray, C. J., & Lopez, A. D. (1996). *Summary: The global burden of disease*. Boston, MA: Harvard University Press.

National Institute for Clinical Excellence. (2004). Depression: Management of Depression in Primary and Secondary Care. London, England: National Institute for Clinical Excellence.

Newman, M. G., Szkodny, L. E., Llera, S. J., & Przeworski, A. (2011). A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: Is human contact necessary for therapeutic efficacy? *Clinical Psychology Review, 31*(1), 89-103. doi: 10.1016/j.cpr.2010.09.008

Nielsen, J. (1993). *Usability engineering*. New York: Academic.

Nielsen, J. (2004). Card sorting: How many users to test. *Evidence-Based User Experience Research, Training, and Consulting*. http://www.nngroup.com/articles/card-sorting-how-many-users-to-test/

Nielsen, J. (2012). Thinking aloud: The #1 usability tool: Nielsen Norman Group: Evidence-Based User Experience Research, Training, and Consulting.

Nielsen, J., & Landauer, T. K. (1993). A mathematical-model of the finding of usability problems. *Human Factors in Computing Systems*, 206-213.

Norman, D. A. (2002). *The design of everyday things*. New York: Basic Books.

Oh, H., Rizo, C., Enkin, M., & Jadad, A. (2005). What is eHealth: A systematic review of published definitions. *Journal of Medical Internet Research, 7*(1).

Olfson, M., Guardino, M., Struening, E., Schneier, F. R., Hellman, F., & Klein, D. F. (2000). Barriers to the treatment of social anxiety. *American Journal of Psychiatry, 157*(4), 521-527.

Pagoto, S. L., Spring, B., Coups, E. J., Mulvaney, S., Coutu, M. F., & Ozakinci, G. (2007). Barriers and facilitators of evidence-based practice perceived by behavioral science health professionals. *Journal of Clinical Psychology, 63*(7), 695-705. doi: 10.1002/jclp.20376

Parush, A., & Yuviler-Gavish, N. (2004). Web navigation structures in cellular phones: The depth/breadth trade-off issue. *International Journal of Human-Computer Studies, 60*(5), 753-770.

Paykel, E. S. (2008). Partial remission, residual symptoms, and relapse in depression. *Dialogues in Clinical Neuroscience, 10*(4), 431-437.

Payne, H. E., Lister, C., West, J. H., & Bernhardt, J. M. (2015). Behavioral functionality of mobile apps in health interventions: A systematic review of the literature. *JMIR mHealth uHealth, 3*(1), e20. doi: 10.2196/mhealth.3335

Pew Research Center. (2014). Mobile technology fact sheet *Fact Sheets*: Pew Research Center: Internet, Science & Tech.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into practice, 41*(4), 219-225.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., . . . Shen, S. (2011). The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry, 168*, 1266-1277.

Possemato, K., Kuhn, E., Johnson, E., Hoffman, J. E., Owen, J. E., Kanuri, N., . . . Brooks, E. (2016). Using PTSD Coach in primary care with and without clinician support: A pilot randomized controlled trial. *General Hospital Psychiatry, 38*, 94-98.

Pratt, L. A., & Brody, D. J. (2014). Depression in the U. S. household population, 2009-2012 *NCHS data brief* (Vol. 172). Hyattsville, MD: National Center for Health Statistics.

Price, M., Yuen, E. K., Goetter, E. M., Herbert, J. D., Forman, E. M., Acierno, R., & Ruggiero, K. J. (2014). mHealth: A mechanism to deliver more accessible, more effective mental health care. *Clin Psychol Psychother, 21*(5), 427-436. doi: 10.1002/cpp.1855

Priest, R. G., Vize, C., Roberts, A., Roberts, M., & Tylee, A. (1996). Lay people's attitudes to treatment of depression: Results of opinion poll for Defeat Depression Campaign just before its launch. *BMJ, 313*(7061), 858-859.

Proudfoot, J. (2013). The future is in our hands: The role of mobile phones in the prevention and management of mental disorders. *Australian and New Zealand Journal of Psychiatry, 47*(2), 111-113.

Rainie, L, & Cohn, D. (2014). Census: Computer ownership, internet connection varies widely across U.S. *FactTank: News in the numbers*. Washington, DC: Pew Research Center.

Rees, C. S., McEvoy, P., & Nathan, P. R. (2005). Relationship between homework completion and outcome in cognitive behaviour therapy. *Cognitive Behaviour Therapy, 34*(4), 242-247.

Renton, T., Tang, H., Ennis, N., Cusimano, M. D., Bhalerao, S., Schweizer, T. A., & Topolovec-Vranic, J. (2014). Web-based intervention programs for depression: A scoping review and evaluation. *Journal of Medical Internet Research, 16*(9), e209. doi: 10.2196/jmir.3147

Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clinical Psychology Review, 32*(4), 329-342. doi: 10.1016/j.cpr.2012.02.004

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task? *Translational Behavioral Medicine, 1*(1), 53-71. doi: 10.1007/s13142-011-0021-7

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

Rosson, M. B., & Carroll, J. M. (2001). *Usability engineering: Scenario-based development of human-computer interaction*. San Francisco, CA: Morgan Kaufmann Publishers.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet, 365*(9453), 82-93. doi: 10.1016/S0140-6736(04)17670-8

Rucci, P., Gherardi, S., Tansella, M., Piccinelli, M., Berardi, D., Bisoffi, G., . . . Pini, S. (2003). Subthreshold psychiatric disorders in primary care: Prevalence and associated characteristics. *Journal of Affective Disorders, 76*(1-3), 171-181.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., . . . Keller, M. B. (2003). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54*(5), 573-583.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research, 17*(7).

Sahota, D. (2014). Android domination to continue in 2014; iPhone loses ground. Retrieved March 15, 2014, from http://www.telecoms.com/210391/android-domination-to-continue-in-2014-iphone-loses-ground/

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine, 33*(2), 155-161. doi: 10.1016/j.amepre.2007.04.007

Sartorius, N., Üstün, T. B., Lecrubier, Y., & Wittchen, H.-U. (1996). Depression comorbid with anxiety: Results from the WHO study on "Psychological disorders in primary health care". *The British Journal of Psychiatry*.

Sauer, J., & Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Appl Ergon, 40*(4), 670-677.

Sauro, J. (2010). A brief history of the magic number 5 in usability testing. from http://www.measuringu.com/blog/five-history.php

Sauro, J. (2012a). Comparison of usability testing methods. *MeasuringU: Usability, customer experience & statistics.* 2015

Sauro, J. (2012b). Measuring errors in the user experience. *MeasuringU: Usability, customer experience & statistics.* 2015

Sauro, J., & Kindlund, E. (2005). *A method to standardize usability metrics into a single score.* Paper presented at the SIGCHI Conference on Human factors in Computing Systems.

Schnall, R., Mosley, J. P., Iribarren, S. J., Bakken, S., Carballo-Diéguez, A., & Brown, W. (2015). Comparison of a user-centered design, self-management app to existing mHealth apps for persons living with HIV. *JMIR mHealth uHealth, 3*(3), e91.

Schueller, S. M., Munoz, R. F., & Mohr, D. C. (2013). Realizing the potential of behavioral intervention technologies. *Current Directions in Psychological Science, 22*, 478-483.

Schueller, S. M., Tomasino, K. N., & Mohr, D. C. (In Press). Integrating human support into behavioral intervention technologies: The efficiency model of support. *Clinical Psychology: Science and Practice*.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Janavs, J., Weiller, E., Keskiner, A., . . . Dunbar, G. C. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry, 12*(5), 232-241.

Shen, N., Levitan, M. J., Johnson, A., Bender, J. L., Hamilton-Page, M., Jadad, A. A., & Wiljer, D. (2015). Finding a depression app: A review and content analysis of the depression app marketplace. *JMIR mHealth uHealth, 3*(1), e16. doi: 10.2196/mhealth.3713

Silva, B. M., Rodrigues, J. J., Canelo, F., Lopes, I. C., & Zhou, L. (2013). A data encryption solution for mobile health apps in cooperation environments. *Journal of Medical Internet Research, 15*(4), e66. doi: 10.2196/jmir.2498

Smith, A. (2015). U.S. smartphone use in 2015: Pew Research Center.

Snyder, C. (2006). *Bias in usability testing*. Paper presented at the Boston Mini-UPA Conference, Natick, MA.

Sorenson, R. L., Gorsuch, R. L., & Mintz, J. (1985). Moving targets: Patients' changing complaints during psychotherapy. *Journal of Consulting and Clinical Psychology, 53*(1), 49.

Spek, V., Cuijpers, P., Nyklicek, I., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological Medicine, 37*(3), 319-328. doi: 10.1017/S0033291706008944

Stowers, K. (2015). *Communication between human factors psychologists and engineers: Challenges and solutions.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Sucala, M., Schnur, J. B., Glazier, K., Miller, S. J., Green, J. P., & Montgomery, G. H. (2013). Hypnosis--there's an app for that: A systematic review of hypnosis apps. *International Journal of Clinical and Experimental Hypnosis, 61*(4), 463-474. doi: 10.1080/00207144.2013.810482

Sunyaev, A., Dehling, T., Taylor, P. L., & Mandl, K. D. (2015). Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association, 22*(e1), e28-33. doi: 10.1136/amiajnl-2013-002605

Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67*(6), 894-904.

Tang, T. Z., DeRubeis, R. J., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 73*(1), 168-172. doi: 10.1037/0022-006X.73.1.168

The MacArthur Foundation Initiative on Depression and Primary Care. (2004). The MacArthur Initiative on depression and primary care at Dartmouth and Duke: Depression management toolkit. Hanover, NH: Dartmouth.

Thomas, K. C., Ellis, A. R., Konrad, T. R., Holzer, C. E., & Morrissey, J. P. (2009). County-level estimates of mental health professional shortage in the United States. *Psychiatric Services, 60*(10), 1323-1328. doi: 10.1176/appi.ps.60.10.1323

Torous, J., Friedman, R., & Keshavan, M. (2014). Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth uHealth, 2*(1), e2.

Trivedi, M. H., Rush, A. J., Ibrahim, H. M., Carmody, T. J., Biggs, M. M., Suppes, T., . . . Kashner, T. M. (2004). The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychological Medicine, 34*(1), 73-82.

Trochim, W. (2010). *Translation won't happen without dissemination and implementation: Some measurement and evaluation issues*. Paper presented at the 3rd Annual Conference on the Science of Dissemination and Implementation, Bethesda, MA.

Tullis, T., & Albert, B. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Burlington, MA: Morgan Kaufmann Publishers.

Usability.gov.). Card sorting. *How To & Tools.* Retrieved April 21, 2015

Usability.gov.). Usability testing. *Methods.* Retrieved January 28, 2016

Van Spall, H. G., Toren, A., Kiss, A., & Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA, 297*(11), 1233-1240. doi: 10.1001/jama.297.11.1233

Videbech, P., & Ravnkilde, B. (2004). Hippocampal volume and depression: A meta-analysis of MRI studies. *The American Journal of Psychiatry, 161*(11), 1957-1966. doi: 10.1176/appi.ajp.161.11.1957

Virzi, R. A. (1992). Refining the test phase of usability evaluation - How many subjects is enough. *Human Factors, 34*(4), 457-468.

Vittengl, J. R., Clark, L. A., & Jarrett, R. B. (2005). Validity of sudden gains in acute phase treatment of depression. *Journal of Consulting and Clinical Psychology, 73*(1), 173-182. doi: 10.1037/0022-006X.73.1.173

Voncken-Brewster, V., Moser, A., van der Weijden, T., Nagykaldi, Z., de Vries, H., & Tange, H. (2013). Usability evaluation of an online, tailored self-management intervention for chronic obstructive pulmonary disease patients incorporating behavior change techniques. *JMIR Res Protoc, 2*(1), e3. doi: 10.2196/resprot.2246

Wallace, A. E., Weeks, W. B., Wang, S., Lee, A. F., & Kazis, L. E. (2006). Rural and urban disparities in health-related quality of life among veterans with psychiatric disorders. *Psychiatric Services, 57*(6), 851-856. doi: 10.1176/appi.ps.57.6.851

Watts, S., Mackenzie, A., Thomas, C., Griskaitis, A., Mewton, L., Williams, A., & Andrews, G. (2013). CBT for depression: A pilot RCT comparing mobile phone vs. computer. *BMC Psychiatry, 13*(1), 49. doi: 10.1186/1471-244X-13-49

Weil, T. P. (2015). Insufficient dollars and qualified personnel to meet United States mental health needs. *The Journal of Nervous and Mental Disease, 203*(4), 233-240. doi: 10.1097/NMD.0000000000000271

Wells, K., Miranda, J., Bauer, M., Bruce, M., Durham, M., Escobar, J., . . . Unutzer, J. (2002). Overcoming barriers to reducing the burden of affective disorders. *Biological Psychiatry, 52*(6), 655-675.

Westfall, J. M., Mold, J., & Fagnan, L. (2007). Practice-based research—"blue highways" on the NIH roadmap. *JAMA, 297*(4), 403-406. doi: 10.1001/jama.297.4.403

Whooley, M. A., & Simon, G. E. (2000). Managing depression in medical outpatients. *New England Journal of Medicine, 343*(26), 1942-1950.

Whooley, M. A., Stone, B., & Soghikian, K. (2000). Randomized trial of case-finding for depression in elderly primary care patients. *Journal of General Internal Medicine, 15*(5), 293-300.

Williams, J. W., Noël, P., Cordes, J. A., Ramirez, G., & Pignone, M. (2002). Is this patient clinically depressed? *Journal of the American Medical Association, 287*(9), 1160-1170. doi: 10.1001/jama.287.9.1160

Wilson, G. Terence. (1998). The clinical utility of randomized controlled trials. *International Journal of Eating Disorders, 24*, 13-29.

Wood, J. R., & Wood, L. E. (2008). Card sorting: Current practices and beyond. *Journal of Usability Studies, 4*(1), 1-6.

Woolrych, A., & Cockton, G. (2001). *Why and when five test users aren't enough.* Paper presented at the Proceedings of IHM-HCI 2001 conference.

Wootten, A. C., Abbott, J. A. M., Chisholm, K., Austin, D. W., Klein, B., McCabe, M., . . . Costello, A. J. (2014). Development, feasibility and usability of an online psychological intervention for men with prostate cancer: My Road Ahead. *Internet Interventions, 1*(4), 188-195. doi: http://dx.doi.org/10.1016/j.invent.2014.10.001

Wright, J. H., Wright, A. S., Salmon, P., Beck, A. T., Kuykendall, J., Goldsmith, L. J., & Zickel, M. B. (2002). Development and initial testing of a multimedia program for computer-assisted cognitive therapy. *American Journal of Psychotherapy, 56*(1), 76-86.

Yang, Y. T., & Silverman, R. D. (2014). Mobile health applications: The patchwork of legal and liability issues suggests strategies to improve oversight. *Health Aff (Millwood), 33*(2), 222-227. doi: 10.1377/hlthaff.2013.0958

Zhang, D., & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International Journal of Human-Computer Interaction, 18*(3), 293-308.

Ziefle, M. (2002). The influence of user expertise and phone complexity on performance, ease of use and learnability of different mobile phones. *Behaviour & Information Technology, 21*(5), 303-311.

Zung, W. W., Broadhead, W. E., & Roth, M. E. (1993). Prevalence of depressive symptoms in primary care. *The Journal of Famly Practice, 37*(4), 337-344.