Embodied Conversational Agents in Internet-Based Cognitive Behavioral Therapy for Depression

Provoost, Simon Jaak

2021

Embodied Conversational Agents in Internet-Based Cognitive Behavioral Therapy for Depression

Bridging the Gap Between Unguided and Guided Interventions

Simon Provoost

Doctoral Committee: prof. dr. Sander Koole (chair)
prof. dr. Rosa Baños
prof. dr. Claudi Bockting
dr. Willem-Paul Brinkman
prof. dr. Mark Hoogendoorn
dr. Saskia Kelders

This thesis was prepared at the Faculty of Behavioural and Movement Sciences of the Vrije Universiteit Amsterdam.

Cover illustration: inimalGraphic

Graphic design and printing: ProefschriftMaken || www.proefschriftmaken.nl

Please cite as:

Provoost, S. (2021). Embodied Conversational Agents in Internet-Based Cognitive Behavioral Therapy for Depression: Bridging the Gap Between Unguided and Guided Interventions. Amsterdam: Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam.

ISBN: 978-94-6423-225-7

VRIJE UNIVERSITEIT

Embodied Conversational Agents in Internet-Based Cognitive Behavioral Therapy for Depression: Bridging the Gap Between Unguided and Guided Interventions

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op dinsdag 1 juni 2021 om 9.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Simon Jaak Provoost

geboren te Oostburg

promotoren: prof.dr. M. M. Riper
prof.dr. T. Bosse

copromotoren: dr. J.J. Ruwaard†

Table of Contents

Dedicated to Jeroen Ruwaard and Johanna van Hoeve-Cijsouw.

May you rest in peace.

CHAPTER 1

1

General Introduction

1 Introduction

Unguided internet-based cognitive behavioral therapy (iCBT) is effective for treating mild to moderate depression, and for many people, can be a good alternative to brief face-to-face cognitive behavioral therapy. However, a common problem associated with unguided iCBT is relatively low adherence to these types of interventions. Adherence is thought to improve when iCBT includes forms of human support, but such support is not always available and can reduce the accessibility of interventions when it is. In the chapters that follow, this thesis explores how support may be delivered automatically by Embodied Conversational Agents, lifelike virtual characters that are capable of emulating humanlike behaviors, and whether such support can positively affect adherence to iCBT.

1.1 Cognitive-behavioral therapy for depression

The World Health Organization estimates that at any given time around 300 million people live with depression worldwide (World Health Organization, 2018), and by 2030 it is expected to be the mental disorder with the largest disease burden in terms of Disability-Adjusted Life Years (Mathers and Loncar, 2006). There are a number of effective treatments for depression (Cuijpers and Dekker, 2005; Cuijpers et al., 2008), one of the most known being cognitive-behavioral therapy (CBT). CBT for depression, which is also used in the treatment of other mental health conditions, such as anxiety (e.g., (Romijn et al., 2019)) and substance use disorders (e.g., (Carroll and Kiluk, 2017)), focuses on identifying and changing the dysfunctional cognitions (e.g., negative thoughts) (Beck et al., 1979) or behaviors (e.g., inactive lifestyle) (Lewinsohn et al., 1976) that caused or help sustain symptoms such as low mood .

In the context of depression, CBT is the most widely researched form of psychological treatment. It is considered more effective than receiving no treatment, about as effective as other forms of psychological therapy (Cuijpers et al., 2013a), and has been shown to be more cost-effective than receiving treatment as usual (Myhr and Payne, 2006). Not everyone responds equally well to CBT however, with factors such as comorbidities (Lopez and Basco, 2015), symptom severity (Cuijpers et al., 2014), and patient attitudes (Renaud et al., 2014) considered to be of influence. Around three in four patients prefer psychotherapy over taking anti-depressants (McHugh et al., 2013), and CBT usually has a well-defined and relatively short duration of around 4 to 6 sessions for mild symptoms, or 10 to 15 sessions for more severe symptoms treated in specialized mental healthcare. This well-defined time-span can be an advantage compared to other treatments such as pharmacotherapy or psycho-analytic therapy, where treatments can last for years.

1.2 Internet-based CBT

Although CBT is effective, there are various reasons why face-to-face CBT is not always available or accessible to patients, such as perceived high costs of face-to-face consultations, stigmatization of people with mental health issues, and a shortage of mental health professionals or their absence in rural areas, that lead to long waiting lists in specialized mental health services (Mohr et al., 2006; Griffiths and Christensen, 2007). This has stimulated the development of internet-based cognitive behavioral therapy (iCBT) and other digitalized psychological interventions (Cuijpers, 1997; Gega et al., 2004; Riper et al., 2010), such that treatment can be offered online. Because CBT generally uses structured protocols, such that the different steps are known beforehand, it is relatively attractive to be transformed to a computerized format. Combined with the strong evidence for its effectiveness in face-to-face therapy settings, this helped make iCBT the most well-researched depression treatment offered via the internet as well (Cuijpers et al., 2013a). Meta-analyses have found iCBT to be effective in decreasing symptoms of depression (Sztein et al., 2018), and, although studies making a direct comparison are limited, about as effective as traditional face-to-face CBT (Andrews et al., 2018).

1.3 Human support

We can distinguish among four types of iCBT based on the kind of support that is part of interventions. The first one is unguided self-help, in which no human support is provided to clients. The second, most common kind of iCBT, is guided self-help. Provided by licensed therapists or, for example, trained volunteers, this type of support focuses on helping clients to work through the self-help material. With therapist time deliberately limited to, for example, one to two hours for an entire treatment, and an emphasis on increasing client motivation, this type of support has also been referred to as 'coaching' support (Mohr et al., 2011). It typically includes elements like feedback on exercises, positive reinforcement, motivational messages, answers to questions about intervention content, and technical support (Mohr et al., 2011; Schueller et al., 2017; Mol et al., 2018). The third and most intense form of online support, both in terms of therapist time and psychotherapeutic content, can be found in online psychotherapy. Client and therapist do not meet face-to-face, and the entire treatment takes place online. The role of the therapist is like the one in face-to-face CBT, not only providing 'coaching' support, but also explaining and introducing each step of the treatment, and providing tailored feedback of a psychotherapeutic nature (e.g., (Ruwaard et al., 2009)). More recently, treatment protocols have been developed in which face-to-face CBT and iCBT are integrated. In this so-called blended therapy, support is provided by the same therapist whom the client speaks to face-to-face (e.g., (Kooistra et al., 2014)). With most of the psychotherapeutic content being discussed in person, the nature and intensity of this fourth type of support is somewhere between that of guided self-help and online psychotherapy.

1.4 Adherence

Unguided self-help interventions generally have relatively small effect sizes and high drop-out rates (Karyotaki et al., 2017), but as they do not require time from therapists or other supportive people, they are easily scalable at relatively low cost. This makes them suitable for preventive purposes, as despite their small effect sizes, when applied to a large population the health gains can be substantial. Effect sizes are found to be larger when iCBT is accompanied by a form of human support (Johansson and Andersson, 2012; Richards and Richardson, 2012), although this can vary depending on the nature of the support (Mohr et al., 2011), symptom severity (Karyotaki et al., 2021) or patient characteristics (Karyotaki et al., 2015). A common observation in guided iCBT interventions is that attrition rates are lower compared with unguided interventions (Spek et al., 2007; Richards and Richardson, 2012), i.e. people are less adherent (non-usage attrition) in unguided interventions, and they tend to drop out (stop using the intervention) more often. In internet interventions, adherence is a measure of how much people experience or are exposed to the content of an intervention, and it can be operationalized and measured in many ways, with more or less precision. Examples are time spent online, number of logins, number of modules completed, and number of exercises completed. A combination of these operationalizations as 'the extent to which individuals are exposed to the content of an intervention' (Christensen et al., 2009) may give the most complete picture since, for example, people may rush through all modules in one go, or log in many times without making any progress.

From a theoretical perspective the precise relationship between human support, adherence, and effectiveness in iCBT is still a topic of discussion. One possibility is offered by the model of supportive accountability, developed by Mohr and colleagues (Mohr et al., 2011). This model suggests that appropriate human support creates an expectation in users that they may have to justify their actions or inactions to the supportive person, i.e., that they are accountable to someone. Human support may also be effective by making interventions more engaging and responsive. This was suggested by Burger and colleagues (Burger et al., 2020), who found that the unguided interventions to which guided interventions are compared in comparative studies, rarely offer technological alternatives to human guidance, such as automated support or extra content. Persuasive technology is an example of such an alternative, with dialogue support through automated reminders or a system taking on the role of a coach considered to have a positive effect on adherence as well (Kelders et al., 2012).

Regardless of the ongoing debate about the precise mechanisms, the positive correlations between human guidance, adherence, and effectiveness are well-established in the literature: interventions with guidance have higher adherence rates and effect sizes (Johansson and Andersson, 2012; Richards and Richardson, 2012), and interventions with higher adherence rates are, generally speaking, more effective (Richards and Richardson, 2012; Karyotaki et al., 2021). In the remainder of this thesis the focus will be on adherence, under the simplified assumption that adherence functions as a mediating variable between human or technological support and intervention effectiveness (Figure 1).

Figure 1. Hypothetical model describing how human and technological support may improve intervention effectiveness with adherence as a mediating variable

1.5 Bridging the gap

There are pros and cons to both unguided and guided iCBT interventions. On the one hand human guidance has a clear positive effect on iCBT, but it requires time from trained volunteers or therapists. On the other hand, while considered less effective, unguided interventions are cheaper to upscale, more readily available, provide more anonymity to users, and may be enhanced with automated forms of support. If human support could somehow be automated while leaving the effective working mechanisms intact, this could give us the best of both worlds, namely guided iCBT that does not require human involvement other than that of the client (or unguided iCBT with automated human-like support). Unguided interventions that provide automated feedback, for example, through automated text or email messages with reminders or positive feedback, are a first step in this direction, but to a large extent lack the human aspects that people can relate or feel accountable to. That this is not the end of the story, however, was already shown more than two decades ago, when Nass and colleagues pointed out that humans tend to treat computers much in the same way as they would treat other humans (Nass et al., 1994; Reeves and Nass, 1996). Since then, large advances in technology have only increased the human-likeness in which computers can operate. Drawing from the field of Artificial Intelligence (AI), the chapters that follow show how I have explored, in collaboration with my coauthors, how the gap between unguided and guided interventions may be bridged using what is arguably one of the computer technologies most aspiring to be human-like: the embodied conversational agent (ECA).

1.6 Embodied Conversational Agents

ECAs can be defined as computer programs that satisfy the following three criteria:

1) They have an embodiment, which could be virtual (avatar) or physical (robot).

2) They interact (converse) with a user, using human communication modalities such as natural language, body gestures, or facial expressions.

3) They can act autonomously upon their environment (agency), usually through some form of reasoning that evokes different behavioral responses to different environmental input.

ECAs can be found in many different application domains, and under a plethora of alternative definitions, that usually emphasize a characteristic feature. Some examples are:

Chatbots. Often applied in, for example, automated customer support on websites, with a focus on automated text-based chatting.

Virtual humans. Aiming to be as realistic (close to human) as possible to elicit more realistic user responses, used in, for example, immersive virtual reality training environments.

Intelligent virtual agents. A popular term in the research community, where the focus is on intelligent behavior.

Relational agents. A focus on developing long-term relationships with users through multiple interactive sessions.

With an estimated 161 synonyms (chatbots.org, 2020), these four are just the tip of the iceberg. It is important to realize that even though the definitions can be synonymous, this is not necessarily always the case. For example, a chatbot that has an embodiment could be considered an ECA, but without one it would just be a chatbot, or possibly also an intelligent virtual agent, depending on how the dialogues are generated. If an ECA communicates through gestures or spoken voice, it is not a chatbot. Communicating in a human-like manner, this same ECA could very well be called a virtual human, but if it is embodied as an animal, this would not be the case. To top it off, all of them could be relational agents if one of their main purposes is to develop a relationship with the user. Throughout most of this thesis I will use the term ECA, as it best captures the capabilities that were under study.

With so many synonyms emphasizing different design aspects, it is not surprising that there is also a vast amount of different ECA design configurations that can be considered. Under our definition of ECAs, embodiments can range from static, cartoon-like, 2-dimensional pictures (e.g., (Provoost et al., 2018)) through dynamic, highly realistic, 3-dimensional virtual characters (e.g., (Swartout et al., 2013)), to physical embodiments such as those of humanoid robots (e.g., (Wainer et al., 2014b)). Communication from the user to the ECA can range from mouse-clicks to proceed through a dialogue (e.g., (Bickmore et al., 2010c)) to multi-media input through real-time video and audio recordings (e.g., (Devault et al., 2014)). From ECA to user, communication can range from text messages (e.g., (Kelders, 2015)) to spoken natural language with matching non-verbal behavior, such as gestures and facial expressions (e.g., (Devault et al., 2014)). The diversity is no less when it comes to the agency that drives an ECA's behavior, with the models behind it ranging from decision trees representing click-through dialogues (e.g., (Bosse and Provoost, 2015)), to large architectures combining multiple machine learning models that allow an ECA to respond in real-time, with appropriate verbal and non-verbal behavior, to a user's spoken words in a microphone (e.g., (Devault et al., 2014)).

Because the reasoning capabilities of an ECA are the most abstract, Figure 2 depicts a simplified example ECA architecture that explains it in some more depth. ECAs generally aim to personalize interaction with their users, such that different kinds of input from users receive different, personalized responses from the ECA. Personalization is not only important to make ECA behavior relevant to users, but, in view of this thesis' topic, also to providing iCBT support (Bickmore et al., 2010a; Mohr et al., 2011; Kelders et al., 2012; Scholten et al., 2017; Kocaballi et al., 2019). Examples of user input are selected responses in a dialogue, specific characteristics contained in their tone of voice, or, in the case of iCBT, user input in the website to which an ECA provides feedback. These characteristics can be saved in a so-called user model such that the information can be used at a later point in a conversation or in a future one. A user model could contain any information relevant to the ECA. Considering the three examples just mentioned a user model could contain the answers to questions in a dialogue, parameter values that tune a machine learning algorithm to recognize a specific person's voice, or values provided to Likert-scale items in an iCBT psycho-education module. The specific information from the user model can then be used by generic algorithms, which I will refer to as the ECAs' reasoning engines, to combine it with what is happening in the conversation at the moment, and come to a personalized response. Using this setup, ECAs can use an empathic expression when a user previously indicated feeling depressed, remember how to recognize and henceforth interpret what users with a different dialect say, or provide positive feedback when a person has completed all exercises in an iCBT treatment module.

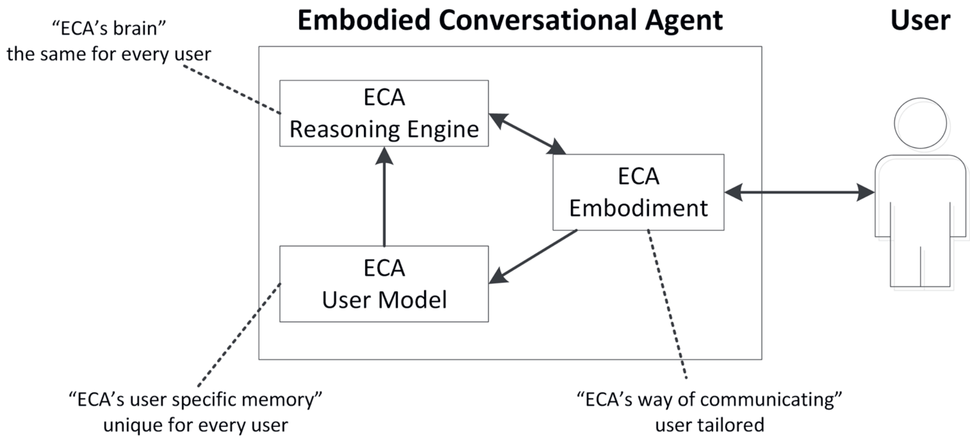

Figure 2. Example of a simplified ECA architecture

Having a specific context significantly reduces the number of considerations that have to be made in designing an ECA, but plenty remain. First it should become clear what kind of role the ECA should fulfil, for example, a coach, therapist or peer user in the context of iCBT. Next, decisions have to be made with regard to the three pillars: embodiment, communication modalities, and reasoning capabilities. There are quite a few recommendations to be found in the ECA literature in this respect, but they are often ambiguous. For example, highly realistic embodiments and communication modalities can be important when immersion and realistic user responses are important, such as in diagnostics (Devault et al., 2014) or social skills training (Bosse and Provoost, 2015) simulations, but it has also been found that high realism sets higher expectations, such that ECAs are not as easily forgiven for mistakes or inconsistencies, resulting in recommendations to focus foremost on proper functionality (van Vugt et al., 2007). Besides what could theoretically be considered the most optimal design configuration, there are also a number of practical issues that come into play. For example, safety concerns can be a barrier to using black-box approaches such as machine learning in the reasoning engine, because such algorithms can be too complex and abstract to evaluate all possible scenarios. Another example relates to the capabilities of software and hardware platforms that the ECA is to work on. These can put further constraints on ECA design. While there are a few ready-made software tools for ECA development, such as the Relational Agent Group's Litebody (Bickmore et al., 2009a) or the University of Southern California's Virtual Human Toolkit (Hartholt et al., 2013), which can be a great asset in developing ECAs, it is not always straightforward to see how ECAs developed in them can be used outside of the software tools. In this case, while the former is hard to integrate in a website and works best as standalone software, the latter requires more computational power than most browsers or old computers could handle. Taking these and other aspects, such as user preferences, integration with an iCBT platform, software and hardware capabilities and limitations, and privacy issues into account, it becomes clear that designing an ECA for iCBT is not a trivial task.

2 Aims of this thesis

Although there were plenty of questions that remained to be answered, it was clear that ECAs can successfully simulate at least some of the human support factors that make guided or face-to-face interventions more effective than unguided interventions. The idea of using ECAs in the context of clinical psychology had already been around for quite some time, with a first serious pitch of the idea dating back to 2010 (Bickmore and Gruber, 2010). Still, we could find little evidence regarding the design and effectiveness of ECAs in clinical psychology to guide their integration with iCBT.

Based on the assumptions that adherence functions as a mediating variable between human support and intervention effectiveness, and that ECAs can simulate human support factors, the primary research question I wanted to answer in this thesis came to be formulated as follows:

Can an ECA that simulates human support factors increase adherence to iCBT?

Based on this primary research question, I set the following two research goals:

(1) Design an ECA for iCBT support based on what is known to work.
(2) Evaluate whether such an ECA can contribute to improving unguided iCBT adherence.

The first goal was of an explorative nature. Without a clear-cut solution at hand we had to determine ourselves how to design an ECA that provides iCBT support. The first five studies described in this thesis aimed to guide this design process.

The second goal was evaluative in nature, questioning whether an ECA providing iCBT support actually improves iCBT results. It was clear that ECA-based solutions were feasible and acceptable, but not whether they could actually improve psychological interventions in a clinically meaningful way. To answer the second question, I designed and developed an intervention with an embedded virtual coach, and prepared everything for evaluation in a pilot randomized controlled trial (RCT).

3 Overview of chapters

This thesis has been written to obtain a PhD by publication. This means that it is composed of a set of separate chapters that, besides this introduction, and the general discussion and summary in the end, have each been submitted to scientific journals as independent papers, with five of them (Provoost et al., 2017, 2018, 2019, 2020; Mol et al., 2018) already published and one submitted at the time of writing. The remainder of this thesis is structured as follows. Chapter 2 reports on the results of a scoping review of ECAs in clinical psychology, in which we aimed to build an overview of the current approaches and evidence base for the use of ECAs in the treatment of common mental health disorders. To gain some experience in using ECAs in our context, and to determine whether a positive effect could be acquired by minimal means, Chapter 3 reports on the results of an experimental study into the effect of very basic virtual agent feedback on adherence to ecological momentary assessment of mood. Chapter 4 describes an experimental study where a sentiment analysis algorithm for the Dutch language, that we were interested in applying in our ECA design, was validated against human judgment. Getting acquainted with the actual content of iCBT support, Chapters 5 and 6 report on the results of mixed-methods studies into iCBT therapist feedback for blended treatments for depression and anxiety disorders respectively. Finally, in Chapter 7 I describe the design and protocol of a pilot RCT of Moodbuster Lite, a four-week iCBT intervention for improving mood with automated virtual coach support. In Chapter 8 the thesis concludes with a general discussion including implications and limitations of this work, as well as directions for future research. In this chapter I will also elaborate on the development process of Moodbuster Lite and the path to ethical approval for the pilot RCT, as to present some further insights in the absence of a report on the study's results. A summary is provided in Chapter 9.

CHAPTER 2

2

Embodied Conversational Agents in Clinical Psychology: A Scoping Review

Simon Provoost, Ho-Ming Lau, Jeroen Ruwaard & Heleen Riper (2017)

Journal of Medical Internet Research, 19(5) e151

doi: 10.2196/jmir.6553

Abstract

Background: Embodied conversational agents (ECAs) are computer-generated characters that simulate key properties of human face-to-face conversation, such as verbal and nonverbal behavior. In Internet-based eHealth interventions, ECAs may be used for the delivery of automated human support factors.

Objective: We aim to provide an overview of the technological and clinical possibilities, as well as the evidence base for ECA applications in clinical psychology, to inform health professionals about the activity in this field of research.

Methods: Given the large variety of applied methodologies, types of applications, and scientific disciplines involved in ECA research, we conducted a systematic scoping review. Scoping reviews aim to map key concepts and types of evidence underlying an area of research, and answer less-specific questions than traditional systematic reviews. Systematic searches for ECA applications in the treatment of mood, anxiety, psychotic, autism spectrum, and substance use disorders were conducted in databases in the fields of psychology and computer science, as well as in interdisciplinary databases. Studies were included if they conveyed primary research findings on an ECA application that targeted one of the disorders. We mapped each study's background information, how the different disorders were addressed, how ECAs and users could interact with one another, methodological aspects, and the study's aims and outcomes.

Results: This study included *N*=54 publications (*N*=49 studies). More than half of the studies (n=26) focused on autism treatment, and ECAs were used most often for social skills training (n=23). Applications ranged from simple reinforcement of social behaviors through emotional expressions to sophisticated multimodal conversational systems. Most applications (n=43) were still in the development and piloting phase, that is, not yet ready for routine practice evaluation or application. Few studies conducted controlled research into clinical effects of ECAs, such as a reduction in symptom severity.

Conclusions: ECAs for mental disorders are emerging. State-of-the-art techniques, involving, for example, communication through natural language or nonverbal behavior, are increasingly being considered and adopted for psychotherapeutic interventions in ECA research with promising results. However, evidence on their clinical application remains scarce. At present, their value to clinical practice lies mostly in the experimental determination of critical human support factors. In the context of using ECAs as an adjunct to existing interventions with the aim of supporting users, important questions remain with regard to the personalization of ECAs' interaction with users, and the optimal timing and manner of providing support. To increase the evidence base with regard to Internet interventions, we

propose an additional focus on low-tech ECA solutions that can be rapidly developed, tested, and applied in routine practice.

1 Introduction

1.1 Background

Internet-based interventions can be effective in the treatment of various mental disorders compared with care-as-usual (e.g., face-to-face treatment) and waiting list control groups (Andrews et al., 2010). Many interventions, especially those aimed at mood, anxiety, and substance use disorders, are based on cognitive behavioral therapy (CBT). These interventions can be unguided or guided, with guidance typically being provided by licensed health professionals or trained volunteers. Guided interventions are typically more clinically effective than unguided ones (Spek et al., 2007; Johansson and Andersson, 2012; Richards and Richardson, 2012). Their superiority is most likely explained by the interaction between the participant and the person providing guidance, and although concepts such as treatment adherence have been suggested as a working mechanism (Mohr et al., 2011), that is, human support having a positive effect on adherence, and in turn on effectiveness, it remains unclear how exactly human support accounts for the difference. Thus, there is good reason to explore whether the gap between guided and unguided interventions can be bridged. In this paper we focus on a potential automated solution: Embodied conversational agents (ECAs).

1.2 Embodied Conversational Agents

ECAs can be defined as "more or less autonomous and intelligent software entities with an embodiment used to communicate with the user" (Isbister and Doyle, 2004). Examples of real-world ECAs are interactive characters in video games and virtual characters that assist customers in Web stores. Conceptually, ECAs consist of three components (Ruttkay et al., 2004). The first is an application interface that allows users to communicate with the ECA and provide it with information. These interfaces can range from Web-based questionnaires to real-time audio and video input. Second, ECAs are endowed with computer models that give them their "mental" capacities, essentially programmed knowledge used to reason over the factual "observations" derived from the interface. Such models can range from concise decision trees in which different answers on a questionnaire lead to different responses by the ECA, to machine learning–based algorithms that classify real-time video and audio input into a user's emotional state, allowing the ECA to react empathically. Third, ECAs have an embodiment, or visual representation, which allows them to communicate with users verbally or nonverbally. Embodiments can range from virtual human characters on computer screens to robots, and communication from text messages to human communication modalities such as speech, gestures, and facial expressions. There are advantages and disadvantages to whatever implementation of the design aspects is chosen. Highly advanced ECAs, for example, those using multimodal and real-time user input such as video recordings and natural language, can be more believable than simplistic ones, but their complexity means that they require more development time, greater technological expertise, and that mistakes (e.g., in interpreting semantics of natural language) become more likely. Low-tech approaches based, for example, on decision tree mechanisms or relatively simplistic graphics can be utilized to deal with these problems, but they also make for a less realistic experience. These kinds of trade-offs make finding the optimal configuration in a certain setting a nontrivial task.

1.3 Are ECAs Ready for Clinical Practice?

Working with an existing Web- and mobile-based cognitive behavioral treatment for depression (Warmerdam et al., 2012), our interest lies with techniques that can be applied in clinical practice. Given the many design decisions that can be made with respect to an ECA's configuration, it is not immediately evident what an ECA should look like, and how it should behave in our context, namely as a bridge between guided and unguided interventions in clinical psychology. Paradigms exist that offer concrete design guidelines with respect to some (e.g., (Baylor, 2011)), or many (e.g., (van Vugt et al., 2009)), of the aspects of ECA development. However, their empirical foundations generally rest on outcome measures such as "user satisfaction," "engagement with the ECA," or "intention to use," and the context is not necessarily that of clinical psychology. Although such measures might be indicative of clinical effectiveness, they do not necessarily translate to the clinical outcomes we aim to improve. For example, even though users might be fully satisfied with an ECA, this does not necessarily mean that the average treatment outcome (e.g., a significant reduction on a clinical measure of depression) will improve. Our chances to successfully bridge the gap between guided and unguided interventions will increase if we can determine how we can apply ECA technology with respect to improving clinical outcomes.

The interdisciplinary nature of ECA research makes almost any intervention that includes an ECA inherently complex. The UK Medical Research Council's (MRC) framework for complex interventions (Craig et al., 2008) defines four phases through which such interventions move before being fully embedded in practice: development, piloting, evaluation, and implementation. The difference between the piloting and evaluation phase is crucial. Whereas interventions might still be subject to changes in the piloting phase, the evaluation phase is characterized by a focus on clinical outcomes and a more rigorous study design. Although routine practice sometimes evolves along different lines, as a golden rule, it is only once an intervention has successfully moved through the evaluation phase and can be considered effective and safe to use, that it becomes of practical value to psychologists.

1.4 Scoping Review

As a first step toward designing our own ECA, we wanted to review the relevant literature in a systematic manner to find out how ECAs had previously been used in psychotherapy, and to what extent the approaches taken were supported by evidence. Initial exploration of the literature to determine a suitable review method revealed a large variety of ECA applications, study designs, and outcome measures, such that a traditional systematic review (more emphasis on hard evidence) or meta-analytic approach (requires comparable outcomes) appeared inappropriate. We therefore adopted the scoping review method (Arksey and O'Malley, 2005). Scoping reviews aim to map the key concepts underpinning a research area, as well as the main sources and types of evidence that are available (Mays et al., 2001). Compared with traditional systematic reviews, scoping reviews address broader topics where many different study designs might be applicable, and do not emphasize quality assessment (e.g., the power of the study or nature of control groups) of the included studies, as the research questions are less specific (Arksey and O'Malley, 2005).

This scoping review aims to inform health professionals about the technological and clinical possibilities and evidence base for ECA applications in clinical psychology, and to provide an overview of the activity in this field of research.

2 Methods

2.1 Study Design

We adopted the Arksey and O'Malley framework for scoping reviews (Arksey and O'Malley, 2005), which distinguishes five different stages: (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results. We took into account recommendations about using an iterative team approach throughout stages (1) to (4) (Levac et al., 2010; Daudt et al., 2013) by having regular discussions with other team members. As theoretical underpinnings to scoping reviews, as well as transparency about the process by which results are obtained, are often lacking (Davis et al., 2009), we also made an attempt to provide clear concept definitions. Stages (1) to (4) are described in this section and stage (5) in the Results section.

2.2 Identifying the Research Question

Given the generic features of human support in psychotherapy, techniques seen in the treatment of disorders other than depression might be applicable in our context as well. Hence, we broadened our scope to include other common mental health disorders known to be a target for e-Mental Health interventions, namely mood disorders, anxiety disorders, post-traumatic stress disorder (PTSD), psychotic disorders, eating disorders, autism spectrum disorders (ASDs), and substance-related disorders.

2.3 Study Identification

Our generic search query was as follows:

embodied conversational agent AND mood disorder OR anxiety disorder OR psychotic disorder OR eating disorder OR autism spectrum disorder OR substance-related disorder

Our list of search terms for the ECA concept included those we observed to be most common, for example, "virtual agent," "virtual character," "virtual human," or "avatar." The terms for the mental disorders were based on those found in PubMed's MeSH (medical subject headings) index. We searched both psychology and computer science databases, including PubMed (psychology), ScienceDirect (interdisciplinary), WebOfScience (interdisciplinary), ACM (Association for Computing Machinery) Digital Library (computer science), and SpringerLink (artificial intelligence). The detailed search strings and search procedures are described in Appendix A1. The final search was conducted in, and included articles published up to, July 2015. References were stored in Microsoft Excel, and duplicates were removed.

2.4 Study Selection

Study selection was conducted by two independent reviewers (SP & HL), who screened titles and abstracts on the sequential eligibility criteria, and then assessed the full-text versions of the remaining articles. A third reviewer (JR) was consulted in case of disagreement. We included full articles that:

(1) were written in English, (2) included an ECA in (3) an applied mental health context, (4) conveyed primary research findings, (5) targeted a mood, anxiety, psychotic, eating, autism spectrum, or substance-related disorder, and (6) described an experimental or focus group study.

Regarding criterion (2), for a software entity to be considered an ECA, it required to, first, have a virtual or physical embodiment (e.g., Figure 1), second, interact with a user and, third, have a reasonable sense of agency, meaning its behavior had to be autonomous, and the software entity had to exhibit some form of reasoning. As for criterion (3), an applied mental health context implied that ECAs were used in an application that aimed to improve patient outcomes directly related to the targeted disorder, or that a reasonable argument could be made that the proposed application could eventually be used to do so.

Figure 1. Examples of embodied conversational agent embodiments: top-left: emotional reinforcement with a smiley face; bottom-left: virtual psychiatric nurse; middle: SPARX's (Smart, Positive, Active, Realistic, X-factor thoughts) guide character; top-right: SimSensei Kiosk virtual counselor; bottom-right: humanoid robot KASPAR

2.5 Charting the Data

Data extraction was conducted independently by two reviewers (SP & HL). Concepts were mapped in four categories: (1) meta-information, (2) study characteristics, (3) study methodology, and (4) ECA characteristics. Precise definitions of the concepts are listed in Appendix A2.

During the data collection process, several concept definitions were refined. In trying to map the studies' intended interventions and provide a taxonomy, we found that from a low level of abstraction, interventions targeted very specific behaviors or skills. Our listing grew so expansive that we considered a higher-level classification to be useful. Something similar could be said for the ECAs' social roles, which were difficult to define precisely given the large variety in applications. In our attempt to provide a useful taxonomy, we tagged all studies with their most predominant social role and intended intervention during discussions with all reviewers present. In these discussions, definitions of the different outcome types and development phases (based on the MRC framework for complex interventions (Craig et al., 2008)) were refined as well, until each of the studies could be tagged unambiguously. The resulting definitions are also listed in Appendix A2.

3 Results

3.1 Main Findings

The search identified N=1117 references. After the removal of duplicates, N=958 references remained. Next, a total of N=862 references were excluded by both reviewers after screening titles and abstracts. Of the remaining N=96 references, the reviewers did not agree about the inclusion based on screening in 78 instances, primarily because it was not entirely clear from the limited information provided by title and abstract whether the ECA inclusion criterion was satisfied (N=40). After full assessment of the 96 remaining articles, disagreement on 8 articles remained and was resolved in a discussion with the third reviewer. Finally, N=54 articles were considered eligible for full review. These 54 articles corresponded to N=49 unique studies (Figure 2). Appendices A3 and A4, respectively, list the included studies' intervention and ECA, and experimental design characteristics. Figure 3 depicts the most important results of the overall summative analysis and Figure 1 gives an illustration of some of the ECAs described in this review, more specifically emotional reinforcement with a smiley face (Agarwal et al., 2013) in the top-left, a virtual psychiatric nurse (Bickmore et al., 2010c) in the bottom-left, SPARX (Smart, Positive, Active, Realistic, X-factor thoughts)'s guide character (Cheek et al., 2014) in the middle, the SimSensei Kiosk virtual counselor (Devault et al., 2014) in the top-right, and humanoid robot KASPAR (Wainer et al., 2014a) in the bottom-right.

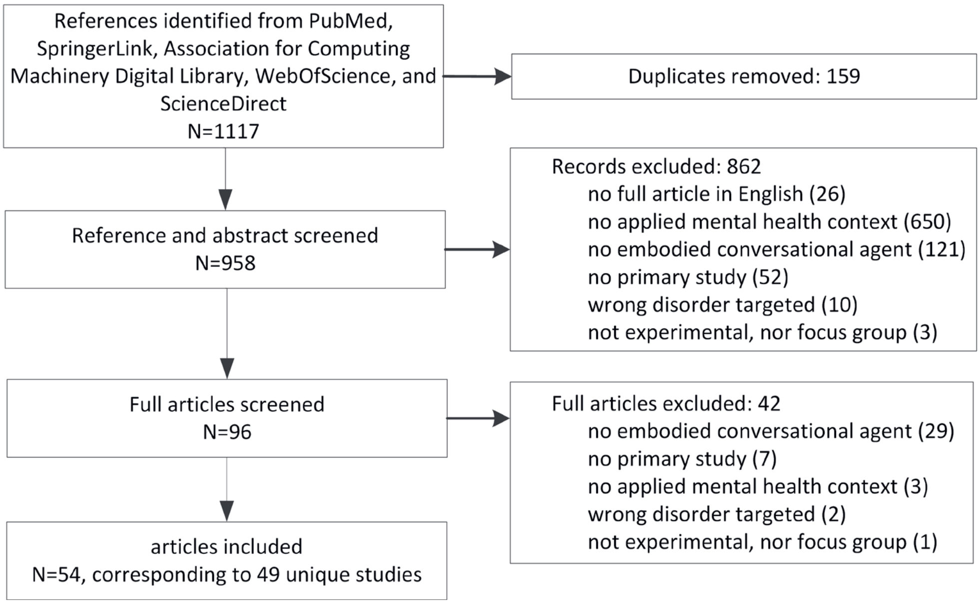

Figure 2. Flowchart describing study identification and selection

3.2 Autism Spectrum Disorders

Over half the studies (n=26) targeted ASDs (Table 1). They either involved a form of social skills training (n=21), aimed at a variety of target behaviors or skills (Figure 3, graph d), or were presented as an educational aid (n=5) to accommodate children with autism's special needs. Autism was the only disorder targeted with robotic applications (n=12), and most of the virtual characters appeared in serious games (n=8). Most ECAs assumed the role of a social interaction partner (n=18) or tutor (n=8), and in two studies (Hopkins et al., 2011; Smith et al., 2014a) social interaction partners were accompanied by a coach who provided additional feedback on their performance. The relative predictability of ECA behavior compared with that of humans, the possibility to repeatedly practice certain behaviors more often than with human partners, and children with autism's fascination for technology were important reasons to explore the use of ECA technology in autism treatment.

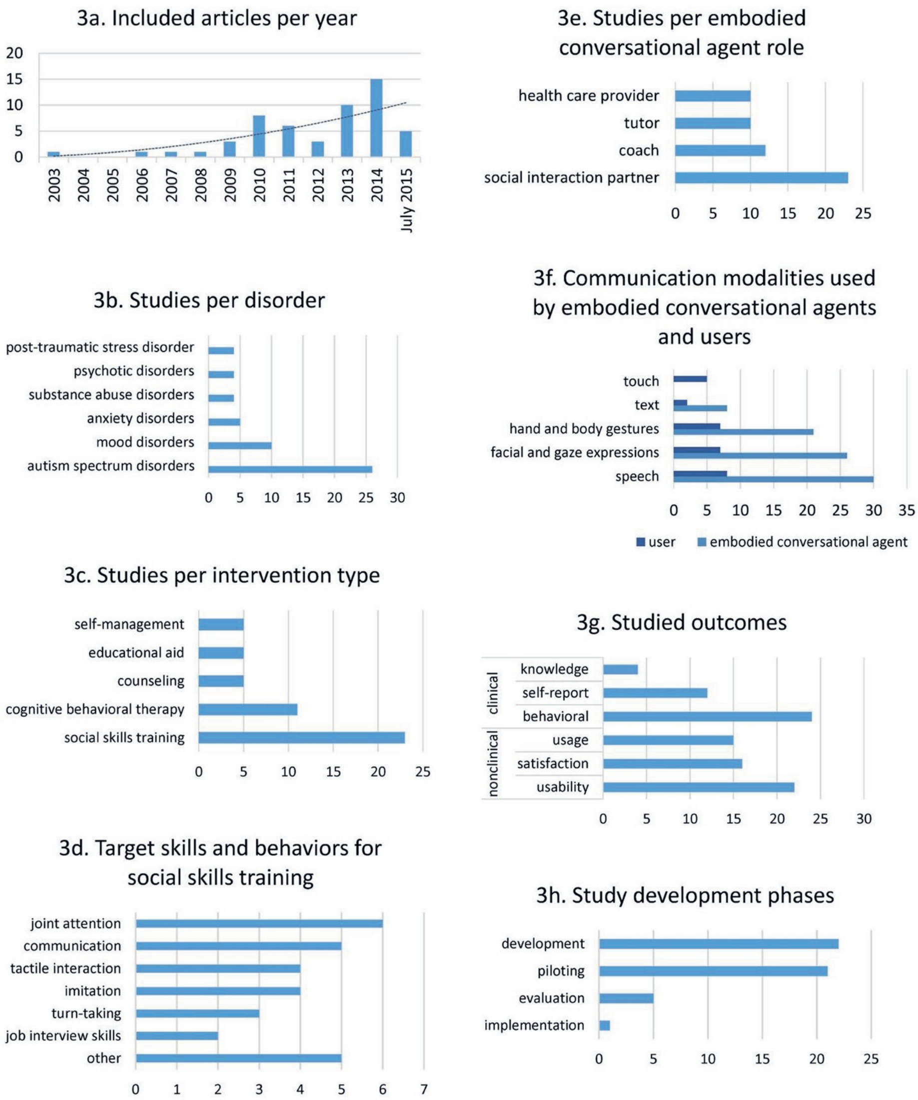

Figure 3. Results of the overall summative evaluation. Note that the categories were not mutually exclusive besides the intervention types and development phases.

3.2.1 Psychotherapeutic Interventions

A first group of ASD studies targeted nonverbal communication skills. In joint-attention skills training, virtual characters (Chen et al., 2010; Alcorn et al., 2011; Hopkins et al., 2011; Agarwal et al., 2013; Bernardini et al., 2014) or a robot (Bekele et al., 2014) would nonverbally cue targets by pointing or gazing at them, after which children were instructed to pay attention to the targets (Chen et al., 2010; Alcorn et al., 2011; Hopkins et al., 2011; Agarwal et al., 2013; Bekele et al., 2014; Bernardini et al., 2014). Imitation skills were taught by asking children to repeat nonverbal gestures made by robots (Boccanfuso and O'Kane, 2010; Fujimoto et al., 2011; Robins and Dautenhahn, 2014; Warren et al., 2015). Other applications focused on tactile interaction by letting children play with robots equipped with tactile sensors (Amirabdollahian et al., 2011; Dickerson et al., 2013; Robins and Dautenhahn, 2014; Costa et al., 2015), turn-taking behavior through playing games with a virtual character (Agarwal et al., 2013) or robot (Boccanfuso and O'Kane, 2010), and facial and emotion recognition by reconstructing faces of dynamic photographs (Hopkins et al., 2011).

Applications focusing on verbal skills used ECAs to teach children communication skills such as general conversation (Lahiri et al., 2011; Tanaka et al., 2015) and antibullying strategies (Milne et al., 2009), and stimulated cooperation in multiplayer games (Jordan et al., 2013; Ribeiro et al., 2014; Wainer et al., 2014a). Lastly, a job interview training application allowed users to practice with a virtual job interviewer (Smith et al., 2014a). It is in this group of ASD studies that we found the only three applications focusing on adults (Jordan et al., 2013; Smith et al., 2014a; Tanaka et al., 2015).

Four of the educational aid applications involved virtual tutors, targeting vocabulary (Cole et al., 2003), daily-life skills (Bamasak et al., 2013), general educational needs of children with ASDs (Konstantinidis et al., 2009), and comprehension of idioms (Shoukry et al., 2015). One study used a humanoid robot to teach children about body postures and aid them in their sense of body consciousness (Palestra et al., 2014).

3.2.2 ECA Technology

From a technological perspective, the focus in many of these studies was on letting ECAs communicate with users through human communication modalities, most notably speech (n=18), facial and gaze expressions (n=15), and hand and body gestures (n=14). A notable exception was touch (n=5), which could only be used by human users.

Few studies employed user models to personalize subsequent interactions. Static user models were mostly used to call the user by name (Konstantinidis et al., 2009; Alcorn et al., 2011; Bekele et al., 2014; Bernardini et al., 2014), and only two studies employed dynamic user models to personalize subsequent interactions. In one study (Chen et al., 2010), emotion recognition was used to structure the narrative of a game, and in another (Smith et al., 2014a) the interactions were structured according to the user's performance and level of rapport with the virtual job interviewer. Only two studies allowed users to enter into a dialog with the ECA, both using menu-based dialogs (Lahiri et al., 2011; Smith et al., 2014a).

3.2.3 Evidence

Although the most commonly studied outcomes were behavioral (n=17), and most of the studies (n=21) were conducted with clinical samples, the sample sizes were generally very small (mean 14.6 (SD 14.2)), and in many cases the behavioral outcomes were short term and restricted to experimental settings (e.g., a researcher's observation of child-robot interaction during an experiment (Robins and Dautenhahn, 2014)). Most studies were still in the development (n=13) and piloting (n=11) phase.

Two ASD studies had moved beyond the development and piloting phase. In the first study (Smith et al., 2014a), the Web platform virtual reality (VR) job interview training was evaluated in a randomized controlled trial (*N*=26). Users were subjected to an interview with a job interviewer, and could ask a coach for feedback regarding their performance. Users could interact with the interviewer through a menu-based system, but they were also given the option of speaking their choices out loud to help them practice their verbal skills. The interviewer could be configured to have different personalities (from friendly and easy-going to mean and asking illegal questions), and based her responses on the user's answers and a user model that kept track of the level of rapport between the two. The interviewer used speech output and conveyed dynamic emotional states through facial and gaze expressions. There was a significant improvement in the user's researcher-scored interview skills in a role-played interview, as well as self-confidence, compared with a control group that received no intervention. Although this result is promising, generalizability to real-world settings and the application's effectiveness compared with conventional job interview skills training programs remain unclear.

The second study used a computer program to improve joint attention, and emotion and face recognition in children with autism (Hopkins et al., 2011). It involved exercises with dynamic realistic photographs of human faces, "coming to life" after successful completion. This intervention also included an animal avatar coach embodied by a realistic photograph that provided additional motivational reinforcement. Following a randomized controlled study (*N*=49), the intervention was found to be more effective in comparison to a control condition in which children used drawing software. Both children with high- and low-functioning autism improved in terms of emotion recognition and observed social skills, while children with high-functioning autism also improved in facial recognition. Similar to the other evaluation study, it remains unclear how the intervention would compare with conventional interventions targeting similar social skills.

3.3 Depression

A total of n=10 studies targeted depression. These studies revolved around CBT interventions (n=4), counseling (n=3), self-management skills (n=2), and social skills training (n=1). Most (n=6) of the applications were Web based, and the social roles fulfilled by ECAs were a coach (n=6) or health care provider (n=5). The anonymity provided by ECAs, their availability compared with humans, their nonjudgmental nature, and the ability for people to practice social interaction in a safe environment were important reasons to explore the use of ECAs in depression treatment.

3.3.1 Psychotherapeutic Interventions

The CBT-based applications targeted symptoms of depression in general, with a virtual coach guiding people with depression through a Web-based intervention (Martínez-Miranda et al., 2014), a photograph of a clinician embodying weekly feedback in a Web-based intervention (Kelders, 2015), and a fantasy character guiding users through a serious game (Cheek et al., 2014). In (Pagliari et al., 2012), requirements for the virtual agent used in (Martínez-Miranda et al., 2014) were determined through a focus group study.

A second group of applications explored elements of counseling with a virtual agent, more specifically the elicitation of self-disclosure, that is, getting people to talk about their problems, by a virtual counselor in an open-ended dialogue (Devault et al., 2014), the elicitation of self-disclosure as well as the provision of relevant information by a Web-based virtual counselor among active soldiers, war veterans, and their families (Swartout et al., 2013), and diagnosis by a virtual therapist guiding users through a Web-based version of the Beck Depression Inventory questionnaire (Pontier and Siddiqui, 2008).

Self-management skills were targeted by an application that supported hospitalized patients during their discharge procedure (Bickmore et al., 2010b), and by a serious game in which people with depression could practice communicating about their health with a virtual doctor (Pinto et al., 2016).

The last study concerned the same job interview training application used in (Smith et al., 2014a), this time targeting people with other psychiatric disabilities, including depression (Smith et al., 2014b).

3.3.2 ECA Technology

Looking at the use of human communication modalities, the most technologically advanced developments take place in counseling interventions from studies conducted by the Institute for Creative Technologies, associated with the University of Southern California (USC-ICT). Over the years, they have developed an extensive framework that allows users to communicate with ECAs in a natural manner through verbal and nonverbal behavior. In one study (Swartout et al., 2013), users could communicate by textual natural language. Using speech and synchronized nonverbal behaviors, the ECA was able to take initiative in the conversation and probe for information related to depression and PTSD. An even more advanced approach in terms of user input was taken in another study (Devault et al., 2014), in which users' speech and nonverbal behavior was taken as input, and used to engage them in open-ended dialogues aimed at self-disclosure about psychological problems. The ECA was endowed with a set of fixed utterances and interview questions, applied back-channel behaviors (e.g., saying "uhuh" and nodding while listening) and empathic responses to build rapport with the user, and used continuation prompts (e.g., a new question) to keep the conversation going.

Another, technologically less advanced, approach that emphasized longer-term user modeling was taken by the Relational Agents Group of Northeastern University. During the past decade, this research group has developed a framework for so-called relational agents that apply a variety of techniques (e.g., daily small-talk, empathic displays, referencing to previous interactions) to develop a long-lasting relationship with the user through menu-based dialogs over multiple interactions (Bickmore et al., 2010b).

A last set of studies focused on modeling users' emotional state in real time based on their interaction with the application itself, for example, based on their answers to depression questionnaires (Pontier and Siddiqui, 2008; Martínez-Miranda et al., 2014).

3.3.3 Evidence

The outcome studied the most (n=8) was user satisfaction. Although the studies focusing on more advanced technologies such as natural interaction and empathy modeling were still in the development (n=5) and piloting (n=2) phase, n=3 studies moved beyond that.

The sole study around an implementation question (Cheek et al., 2014) concerned a focus group study (*N*=16) of the acceptability of SPARX, a gamified CBT intervention developed in New Zealand, in which players can, for example, use a staff to shoot physically manifested negative thoughts. ECAs as such are not a predominant theme in the SPARX game, but regular mention of a guide character is made. Players choose their own avatar that provides instructions throughout the game in dialogs with the user. Australian participants indicated that it was important that the guide's gender could be customized, did not mind its foreign accent, and liked the idea of being able to socialize with it. Being a focus group study, it remains unclear whether these results would hold in an experimental setting.

The first evaluation study (N=134) used the photograph of a clinician as an embodiment to deliver automated motivational support in a computerized acceptance and commitment therapy (Kelders, 2015). Users could not interact with the ECA directly, and personalized support occurred on a predefined schedule through a user model based on the user's actions in the intervention. Participants receiving automated ECA feedback were found not to be significantly less involved than those receiving real human support. Although this result was very interesting in the sense that ECA support was compared with real human support, the ECA itself made little use of state-of-the-art ECA technology, and therefore gives us little to go on in terms of ECA design. The second evaluation study concerned another randomized controlled study (N=37) using the job interview training application (Smith et al., 2014b) that was also used for ASDs (Smith et al., 2014a). Again, the users' researcher-scored interview skills in a role-played interview, as well as users' self-confidence, improved significantly compared with a control group that received no intervention.

The development and piloting studies provided us with some initial evidence that practicing health communication with virtual health care providers in a serious game can be efficacious (Pinto et al., 2016), that ECAs in a CBT-based intervention should have a coaching role, be configurable, adaptable, trustworthy, guiding rather than directive, and capable of empathic expressions without reflecting negative ones back to the user (Pagliari et al., 2012), that ECAs endowed with empathy are more highly valued than those without it (Pontier and Siddiqui, 2008; Martínez-Miranda et al., 2014), that people do not experience less rapport when interacting with an ECA than when interacting with a human (Devault et al., 2014), that people appreciate the anonymous nature of interacting with an ECA (Swartout et al., 2013), and that people with depression experience a stronger working alliance with a virtual nurse guiding a hospital check-out procedure than do the nondepressed (Bickmore et al., 2010b).

3.4 Anxiety Disorders

There were *N*=5 studies that targeted anxiety disorders, either with CBT (n=2) or counseling (n=3) interventions. ECAs assumed the role of a social interaction partner (n=3), a health care provider (n=2), and a tutor (n=1), and most of them were implemented in stand-alone software (n=3). Reasons to use ECAs for anxiety disorders were similar to those mentioned for depression.

3.4.1 Psychotherapeutic Interventions

The counseling studies experimented with various techniques to elicit self-disclosure in counseling sessions with a virtual agent. While we already discussed one study (Devault et al., 2014) for depression, two other studies focused solely on eliciting personal information from people with anxiety in the context of finding a new roommate (Kang and Gratch, 2010) and counseling (Kang and Gratch, 2012). In the CBT-based applications, virtual animals helped children to conquer performance anxiety in a serious game (Schmidt et al., 2013), and a virtual character evoked anxiety in a VR environment (Rinck et al., 2010).

3.4.2 ECA Technology

Most innovative here are the counseling studies, again conducted by USC-ICT. Although all three ECAs are based on the same framework, those described in (Kang and Gratch, 2010) and (Kang and Gratch, 2012) differ from (Devault et al., 2014) in that a so-called "Wizard of Oz" paradigm was applied to control the ECAs' verbal behavior, that is, it was controlled by the researchers. Whereas this violated our definition of agency in terms of verbal behavior, the so-called rapport agents' nonverbal behavior was completely automated. Interpreting the phonetic aspects of a user's speech input, as well as video recordings of his or her nonverbal behavior, they were able to display appropriate nonverbal behaviors themselves.

3.4.3 Evidence

The applications we considered were still in the development (n=2) and piloting (n=3) phase, and there was no predominant outcome measure used. Although some studies worked with large sample sizes (N=351 in (Devault et al., 2014), and N=108 in (Kang and Gratch, 2010)), none of the studies experimented with clinical samples.

In these studies, most relevant to our purpose were the findings that people with elevated levels of social anxiety may find it easier to disclose personal information to an ECA than to a human (Kang and Gratch, 2010), that human backstories may be more effective in this respect than (true) computer backstories (Kang and Gratch, 2012), and that highly anxious people approached a character in a VR environment more slowly, and kept more distance, than less-anxious ones (Rinck et al., 2010).

3.5 Post-Traumatic Stress Disorder

A total of n=4 studies targeted PTSD. Besides two studies on counseling interventions that also targeted depression (Swartout et al., 2013; Devault et al., 2014), one study proposed a virtual coach in a CBT-based Web-based platform (Tielman et al., 2014), and one concerned a virtual guide in a serious gaming healing environment (Morie et al., 2009). The studies involved ECAs in the role of a coach (n=3) and a health care provider (n=1).

3.5.1 Psychotherapeutic Interventions

In the first study (Morie et al., 2009), a fantasy character acted as an engaging information repository in a virtual healing environment for returning soldiers that stimulated social comradery, healing activities, and personal exploration. The other study involved a focus group of experts in trauma treatment, in which design requirements were gathered for a virtual agent supporting a Web-based exposure therapy-based application (Tielman et al., 2014).

3.5.2 ECA Technology

From a technological perspective, the most interesting developments took place in the two studies we already discussed. The guide in (Morie et al., 2009) was implemented as a virtual character in a private space built in the Second Life virtual worlds platform, but the details about its design remained unclear.

3.5.3 Evidence

All studies (n=4) were still in the development phase. Even though (Morie et al., 2009) had an impressive sample size (N=700), the focus was still on usability of the healing environment itself. Some examples of suggested guidelines for the virtual coach from the focus group study (N=10) (Tielman et al., 2014) were that it should acknowledge patients' feelings, remind them of their goals when they indicate they wish to quit, be factual in complimenting, and never express negative emotion.

3.6 Psychotic Disorders

The n=4 studies involving psychotic disorders revolved around social skills training (n=2) and self-management (n=2). Aside from the Web-based job interview training application that also targeted depression (Smith et al., 2014b), two applications were implemented in stand-alone software, and one in a VR environment. Interestingly, this set of studies was the only one to consider ECAs in all four social roles. Important reasons to explore the use of ECAs in the treatment of psychotic disorders were that social skills could be practiced in a safe environment, and that ECAs can always be available to provide support or information.

3.6.1 Psychotherapeutic Interventions

Two studies applied the Relational Agent framework described in the section on depression (Bickmore et al., 2010b), and used an ECA to host a system that provided general lifestyle support with an emphasis on promoting medication adherence for people suffering from schizophrenia over a 1-month period (Bickmore et al., 2010c;

Puskar et al., 2011). In the other study, people with schizophrenia could practice conversational skills with virtual characters in a VR social situation (Ku et al., 2007).

3.6.2 ECA Technology

The self-management interventions that were based on the Relational Agent framework used a similar set of techniques as in (Bickmore et al., 2010b) to develop a long-term relationship between a virtual psychiatric nurse and the user. The conversations in (Ku et al., 2007) followed a branching tree model approach, which allowed users to communicate through a multiple choice menu. A virtual coach could help users in case they ended up in negative situations. All of the studies allowed users to interact with ECAs through menu-based dialogs.

3.6.3 Evidence

Although the studies that have not already been discussed under the various disorders were all in the piloting phase (n=3), and sample sizes were fairly small (mean 19.8 (SD 11.9)), they all studied clinical populations. Whereas usability (n=3) and satisfaction (n=4) outcomes were studied most often, with positive results, there is some initial evidence that the Relational Agent applications helped people with schizophrenia to adhere to their medication intake (Bickmore et al., 2010c; Puskar et al., 2011), and that VR social situations evoke similar negative symptoms in people with schizophrenia to what would be expected from real-world situations (Ku et al., 2007).

3.7 Substance Use

All ECA applications (n=4) targeting substance use were Web based and included CBT elements. ECAs assumed the role of a coach (n=2) and a health care provider (n=2). The main reason to use ECAs in the context of substance use is that they are more available than supportive humans would be, thereby increasing accessibility.

3.7.1 Psychotherapeutic Interventions

In (An et al., 2013), a "makeover host" was used to deliver and highlight personalized content in an intervention called REALU2, targeting healthy lifestyle behavior with an emphasis on smoking cessation. Smoking cessation was also the topic of (Grolleman et al., 2006), which investigated the acceptability of a proposed virtual agent for an intervention based on motivational interviewing. Motivational interviewing was used in a brief intervention targeting problematic drinking behavior in (Lisetti et al., 2013; Yasavur et al., 2014).

3.7.2 ECA Technology

The makeover host in (An et al., 2013) delivered personalized messages based on how users interacted with the intervention, but the details of its design remained unclear. The motivational interview intervention described in (Lisetti et al., 2013) allowed users to interact with a virtual counselor through a menu-based system, and their facial expressions were recorded to deduce their emotional state. The combination of user input and emotional state allowed the ECA to guide the conversation and respond empathically. In (Yasavur et al., 2014), efforts were made to make interaction with the same system more natural by using speech rather than menu-based input.

3.7.3 Evidence

This set of studies contained one study in the evaluation phase, which had the highest number of participants (N=1317 adults with a history of smoking) out of all the studies considered. In a randomized controlled trial, the ECA intervention was found to be more effective in reducing self-reported smoking than a control condition in which participants used an intervention based on general lifestyle support. Including peer support further boosted the ECA intervention's effectiveness. Because the ECA's design was not described in detail, and because the intervention with the ECA was not compared with one without it, it remains unclear how the ECA itself contributed to the results. Some evidence for the importance of empathic behavior by ECAs was provided in a randomized controlled trial on the brief motivational intervention (Lisetti et al., 2013). The ECA using an empathy module performed significantly better than an ECA without it on various outcome measures, but long-term effects regarding substance use remained unclear.

Table 1. Summative results per disorder

	ASD	Depression	Anxiety	PTSD	Psychotic	Substance Use
Total number of studies	**n=26**	**n=10**	**n=5**	**n=4**	**n=4**	**n=4**
Interventions						
social skills training	21	1			2	
educational aid	5					
CBT		4	2	1		4
counseling		3	3	2		
self-management		2		1	2	
Platform						
serious game	8	2	1	1		
stand-alone software	4	2	3	1	2	
robotics	12					
virtual reality	1		1		1	
web-based	1	6		2	1	4
Social Role						
social interaction partner	18	1	3		2	
tutor	8		1		1	
coach	2	6		3	1	2
health-care provider		5	2	1	2	2
ECA Human Communication Modalities						
speech	18	5	3	2	4	2
facial and gaze expressions	15	5	4	2	3	1
hand and body gestures	14	3	3	2	2	1
text	2	3	1	1		1
touch						
User Human Communication Modalities						
speech	3	2	3	1	1	1
facial and gaze expressions	3	1	3	1		1
hand and body gestures	4	1	3	1		
text	1	1		1		
touch	5					

Personalization						
static user model	4					
dynamic user model	2	7	1	2	3	3
menu-based dialog	2	4		1	4	1
natural language dialog		2	1	2		1
Development Phase						
development	13	5	2	4		1
piloting	11	2	3		3	2
evaluation	2	2			1	1
implementation		1				
Outcomes						
usability	11	5	2	3	3	2
satisfaction	3	8	1	2	4	2
usage	11	1	1		2	1
behavioral	17	2	3		3	
self-report	2	4	3	1	2	2
knowledge	4					
Study Participants						
Mean N (SD)	14.6 (14.2)	93.8 (108.8)	107.4 (141.0)	293 (306.7)	19.8 (11.9)	380.5 (624.8)
min N, max N	1, 49	8, 351	15, 351	10, 700	10, 37	35, 1317
clinical sample	21	7		1	4	2
adult sample	5	7	4	4	4	4

4 Discussion

4.1 Principal Findings

This review aimed to inform health professionals about the technological possibilities and evidence base for ECA applications in clinical psychology, and to provide an overview of the activity in this field of research. Research on the use of ECAs in psychotherapy is emerging (Figure 3, graph a), and we reviewed N=49 studies of which the majority targeted ASDs (Figure 3, graph b). A general distinction could be made between applications in which ECAs were used as an adjunct to an intervention that could also have been used independently, and applications in which the interaction between the ECA and user was central. The former were mostly CBT-based programs, educational aids, and self-management interventions, whereas the latter were mostly social interaction skills training interventions and counseling interventions. Social skills training interventions were by far the most popular for ASDs, which also made them the predominant type of intervention overall (Figure 3, graph c). As a result, ECAs in the role of a social interaction partner were the most frequent (Figure 3, graph e). The large variety in ECA applications and types of interventions (e.g., Figure 3, graph d) made it a nontrivial task to provide a taxonomy of interventions and ECA social roles. Although clinical behavioral outcomes were studied most often (Figure 3, graph g), they were in many cases restricted to pre-post measurements within experiments that had relatively small sample sizes. Consequently, few studies exceeded the piloting phase (Figure 3, graph h).

4.1.1 ECA Technology

The balance in terms of the ability to communicate through human communication modalities such as gestures, expressions, and speech highly favors ECAs compared with human users (Figure 3, graph f). Nevertheless, work to shift this balance has been conducted in research on social interaction training and counseling interventions. The latter made a lot of use of technologically more advanced user modeling (e.g., real-time emotional states), and innovative ways for humans to communicate with computer systems, for example, by interpreting natural language, phonetic aspects from speech input, and recorded nonverbal behavior. While ASD and counseling interventions have a more short-term focus, there has also been some activity in establishing longer-term relationships by using less technical user models, the input of which came from dialog-tree systems and indirect communication with ECAs through the intervention interface itself. Given the development phases of the studies in which these technologies were applied, the more high-tech solutions seem to be most suited for experimental research into the client-therapist relationship or screening for disorders such as depression or PTSD (e.g., (Devault et al., 2014) or (Swartout et al., 2013)). At present, the more low-tech approaches have been evaluated most thoroughly and, therefore, seem most promising for direct application in routine clinical practice.

4.1.2 Evidence Base

From the evaluation phase studies we learned that there is reasonable evidence that (1) ECAs can have a positive effect on user engagement and involvement, (2) they can be effective in this sense as an adjunct to already existing interventions for mental disorders, and (3) it is important for them to convey empathy when interacting with users. An important limitation of the evaluation studies we considered is the nature of the control groups. Besides the study by Kelders (Kelders, 2015), none compared an ECA intervention to a conventional treatment used in routine practice, or compared the intervention with the ECA to the intervention without it. Additionally, the more rigorous studies either put little emphasis on the role and design of the ECAs, or targeted disorders indirectly (e.g., job interviews skills). Although most of the developmental and piloting studies show promising results with respect to usability and user acceptance, we still have little hard evidence that the proposed applications are reasonable alternatives to established treatments, or that ECAs used as an adjunct to existing interventions, bridging the gap between guided and unguided interventions, make them more clinically effective.

4.2 Future Work

In Table 1, some notable blank spots can be identified. Most obvious is the scarcity of evaluation and implementation phase studies, which requires more research with larger sample sizes, suitable control groups, and clinical populations. In this respect, the emerging nature of the field is a reassuring consideration. Two other notable blank spots are Web-based ECA applications for ASDs, and the use of ECA technology in VR applications in general. Taking into account the individual studies' descriptions, there is also still quite some research to be done with regard to effective ECA configurations, by comparing different parameter settings in controlled studies.

4.2.1 Web-Based CBT

To revert to delivering support in Internet interventions, we know that not all people require the same amount of support, considering, for example, that people with low intrinsic motivation benefit more from human support than those who are already motivated or prefer to work on their own (Mohr et al., 2011). User models on, for example, patient motivation could contribute to an accurate timing of ECA support, for example, by providing support when motivation is low and not disturbing them when motivation is already high. Considering the working mechanisms through which human support may increase the effectiveness of Internet interventions, little work has been conducted in the area of treatment adherence, making this a pertinent target around which to focus our efforts with respect to increasing motivation.

Another important point related to extending the evidence base is illustrated by the Help4Mood project (Martínez-Miranda et al., 2014) and the study by Kelders (Kelders, 2015). The Help4Mood project attempts to integrate a Web-based CBT-based treatment with a technologically advanced ECA that communicates through speech, and that is endowed with a dynamic emotion model used to convey empathy in real time. In the study by Kelders (Kelders, 2015), automated textual feedback in an already existing intervention was embodied by accompanying it with the photograph of a clinician. Help4Mood is technologically more challenging, but requires a long period of development and piloting. Because the ECA plays such a central role in the intervention, it is far from straightforward to simply “add” the ECA as an adjunct to an already existing and well-evaluated intervention to study its effects. Rather, the intervention would have to be built around the ECA framework, and the resulting new intervention would once again have to proceed through the development phases. This is a general issue whenever we consider adopting one of the other ECA frameworks we came across, such as the Relational Agent Group’s Litebody (Bickmore et al., 2009a), and USC-ICT Virtual Human Toolkit (Hartholt et al., 2013), especially if we consider that the input used in Help4Mood is still relatively straightforward to interpret compared with human speech or nonverbal behavior.

4.2.2 Low-Tech Approach

If we want to investigate how to improve interventions that have already been set out in the field, a “low-tech” approach similar to (Kelders, 2015) can be advocated because it (1) saves development time that can be used to design and set up larger studies, (2) forces us to think about the core attributes that can make the ECA effective, and (3) makes it easier to judge whether it is safe to use the ECA in a clinical setting with real patients. Given the sparse evidence on the clinical effectiveness of ECAs thus far, this approach and its three advantages may be just what we need to study how we can effectively use ECAs in existing Internet interventions: it will be easier to (1) conduct studies that move beyond the piloting phase, (2) identify the core attributes that make ECAs effective in Internet interventions such that ECA design can be more focused and less time-consuming, and (3) conduct experiments with clinical populations such that we can study ECAs’ effects on clinical outcomes.

4.3 Limitations

Our definition of ECAs had three components. With respect to the embodiment and interaction capabilities, we took a liberal stance, but our requirement of agency was rather conservative (autonomous behavior and reasoning). This excluded a fair number of studies (29 during the screening of full articles) from our review that some might consider to be relevant. Regarding the criterion of autonomous behavior, we excluded quite some studies using what is often called a “Wizard of Oz” paradigm, in which the ECA’s behavior is not controlled by a software entity, but by a human operator instead. Examples are a study in which an ECA representing the hallucinated voices of people with schizophrenia spoke the transformed utterances of a therapist

(Leff et al., 2013), and one in which a robot aimed at improving the mood of hospitalized children suffering from cancer was under the control of a researcher (Alemi et al., 2014). An example of a study that was excluded because the embodied characters lacked reasoning capabilities, that is, the ECA would act the same regardless of user input, was (Wallace et al., 2010).

Although ECA research is almost inherently interdisciplinary, we refrained from going too deep into the technological aspects. This was because our target audience consisted of health professionals with a generally less technical background and we wanted to focus on opening up the ECA domain for them as well as providing them with an overview of the available evidence for application in routine clinical practice. For this reason, we refrained from a highly technical discussion of, for example, verbal and nonverbal ECA capabilities. However, it has to be noted that, depending on how one would like to use ECAs in future work, many more detailed questions could be investigated surrounding ECA design aspects, such as the required capabilities for, and their impact on, specific disorders or types of ECA interventions. With respect to our search strategy, we looked specifically for articles that mention ECAs. As exemplified by the sole included article on SPARX, which is actually supported by more research than reviewed here (e.g., (Merry et al., 2012)), there is a possibility that we missed out on articles describing, for example, serious games or VR environments in which ECAs are used, but not specifically mentioned.

Another limitation relates to the bibliographic databases we considered. Computer science research publications are more dispersed than those of psychology research and computer science databases are less suited to systematic searches. Although our interdisciplinary approach was already broader than what is usual in psychology research, there is a possibility that we might have missed relevant research in, for example, the IEEE (Institute of Electrical and Electronics Engineers) Xplore digital library or Google Scholar. We refrained from using IEEE Xplore digital library due to practical constraints and Google Scholar because its search algorithm is often updated and personalized, which makes it difficult to replicate search results. Additionally, due to practical constraints, we did not conduct searches in the gray literature or manual searches through cross-referencing, nor did we conduct a follow-up search after the original one.

While the idea of applying ECAs in psychotherapy is far from new (e.g., (Bickmore and Gruber, 2010; Hudlicka, 2013)), to our knowledge this is the first review to specifically consider psychotherapeutic applications of ECAs in a systematic manner. There are several areas of research that are closely related to ours. One of these is social robotics research, which often focuses on providing company to elderly people (e.g., (Broekens et al., 2009)) or people suffering from dementia (e.g., (Mordoch et al., 2013)). While this area of research could be seen as an attempt to prevent the psychological consequences that might ensue from loneliness, we did not consider the focus to be on psychopathology. Robotic applications targeting autism have previously been reviewed (e.g., (Diehl et al., 2012)), but without the constraints implied by our ECA concept. Consequently, many of these robotic applications do not adhere to our criterion of agency, that is, they do not act autonomously or intelligently. Besides research on robotics, there is a large corpus of literature on the application of virtual agents in other highly relevant domains from which we can draw inspiration. Although this review focuses solely on psychotherapeutic applications, there seems to be little reason not to consider, for example, motivational (Baylor, 2011), pedagogical (Baylor and Kim, 2004), or lifestyle-support agents (Smith et al., 2008).

4.4 Conclusions

Research into the psychotherapeutic application of ECAs is emerging. We identified 49 studies, with over half of them focusing on autism. The field is characterized by a large variety in all its aspects, for example, type of intervention, target behavior, platform, ECA embodiment, communication modalities, ECA “mental” states, and study design. While there are several studies surpassing the development and piloting phases, as might be expected in a relatively new field, evidence about the clinical effectiveness of ECA applications remains sparse. Technologically advanced ECA applications are very interesting and show promising results, but their complex nature makes it difficult for now to prove that they are effective and safe to use in clinical practice. Therefore, at present, clinical practice seems well served by an additional focus on a more low-tech approach based on the elementary principles that make ECAs effective, that can progress through the development and piloting phases at a faster pace, and that can therefore more easily be proven to be safe and effective for routine clinical practice.

CHAPTER 3

3

Mood Mirroring with an Embodied Virtual Agent: A Pilot Study on the Relationship Between Personalized Visual Feedback and Adherence

Simon Provoost, Jeroen Ruwaard, Koen Neijenhuijs, Tibor Bosse & Heleen Riper (2018)

Communications in Computer and Information Science, vol 887
doi: 10.1007/978-3-319-94779-2_3

Abstract

Background: Human support is thought to increase adherence to internet-based interventions for common mental health disorders, but can be costly and reduce treatment accessibility. Embodied virtual agents may be used to deliver automated support, but while many solutions have been shown to be feasible, there is still little controlled research that empirically validates their clinical effectiveness in this context.

Methods: This study uses a controlled and randomized paradigm to investigate whether feedback from an embodied virtual agent can increase adherence. In a three-week ecological momentary assessment smartphone study, 68 participants were asked to report their mood three times a day. An embodied virtual agent could mirror participant-reported mood states when thanking them for their answers. A two-stage randomization into a text and personalized visual feedback group, versus a text-only control group, was applied to control for individual differences (study onset) and feedback history (after two weeks).

Results: Results indicate that while personalized visual feedback did not increase adherence, it did manage to keep adherence constant over a three-week period, whereas fluctuations in adherence could be observed in the text-only control group.

Conclusions: Although this was a pilot study, and its results should be interpreted with some caution, this paper shows how virtual agent feedback may have a stabilizing effect on adherence, how controlled experiments on the relationship between virtual agent support and clinically relevant measures such as adherence can be conducted, and how results may be analyzed.

1 Introduction

Internet-based psychotherapeutic interventions, also referred to as eMental Health interventions, can be effective in the treatment of various mental disorders when compared to face-to-face interventions (Andersson et al., 2014). Many interventions that target common mental health disorders such as mood, anxiety, and substance use disorders, are based on cognitive behavioral therapy (CBT). Internet-based CBT interventions are either guided, or self-guided, with guidance usually being provided by health professionals or trained volunteers. It has been found that guided interventions are generally more clinically effective, e.g., reductions in symptomatology, than unguided interventions (Richards and Richardson, 2012). While the precise contribution of human support remains unclear, a number of working mechanisms have been suggested (Schueller et al., 2017). One such mechanism is that human support contributes to patients' motivation to complete an intervention, which in turn may increase adherence (Mohr et al., 2011). Indeed, it has been shown that adherence may be superior when human support is available (van Ballegooijen et al., 2014), and that non-optimal exposure or non-adherence to interventions, e.g., not completing exercises or dropping out of interventions early, tends to reduce their clinical effectiveness (Donkin et al., 2011).

The study described in this article is part of a project in which we are looking to bridge the gap between guided and unguided internet-based CBT interventions by automating support through the use of embodied virtual agents. From a literature review of their application in the treatment of common mental health disorders, we concluded that few studies have explored their use in a supportive role to online CBT-based interventions. Although a number of applications seemed feasible and promising, there is still little evidence for their impact on clinically relevant outcomes such as symptom reduction or adherence (Provoost et al., 2017). Clinical psychology is an applied science, however, which means that there is a strong emphasis on empirical validation when introducing novel technologies. Although from a technological perspective, lots of interesting solutions have been, and are being developed, they cannot be applied in clinical practice without such validation. The present study represents the first in a series of controlled studies in which we aim to discover how and whether virtual agent support can contribute to eMental Health interventions' clinical effectiveness.

Because a detailed study of clinical outcomes in controlled settings, such as symptom severity, requires clinical study populations (ethical implications) and follow-up measurements (long timespan), we chose to study adherence as an outcome measure in this pilot study, on the assumption that it is a potential mediator for clinical effectiveness. We opted for Ecological Momentary Assessment (EMA) as an intervention strategy, also referred to as experience sampling, which refers to the repeated sampling of subjects' current behaviors and experiences in real time, in subjects' natural environments (Shiffman et al., 2008). EMA can be a component of internet-based CBT interventions, for example, to measure fluctuations in mood (Wenze and Miller, 2010). Additionally, it is a clear measure of adherence, as patients either do or do not respond to EMA requests.

In the remainder of this paper we describe our explorative pilot study, in which we test the hypothesis that virtual agent support can increase adherence to EMA requests. With respect to the virtual agent design, we opted for a simple approach that fulfills the ECA criteria (the agent has an embodiment, communicates with the user, and uses a form of reasoning to simulate agency (Isbister and Doyle, 2004)) in a minimalistic manner. Although frameworks exist for the development of ECAs with their full range of verbal and non-verbal capabilities (e.g., (Bickmore et al., 2009a; Gratch and Hartholt, 2013)), using them for the development of virtual agents, and their subsequent integration with existing EMA platforms, was considered too time-consuming for this study. Moreover, agents do not necessarily need to be very complicated for motivational purposes, because it has been shown that even the mere presence of an embodied agent can improve user motivation, for example, when shown next to a chat dialog box in which instructions and feedback for an interactive game are displayed (Mumm and Mutlu, 2011). A more detailed description of the experiment and the agent's design is provided in the Methods section.

2 Methods

2.1 Design

We conducted an explorative controlled pilot study with a two-stage randomized between-subject design. Participants self-monitored their mood on a smartphone EMA application in which a virtual character could give personalized visual feedback to user responses by mirroring their reported mood state. Approval for the study was obtained from the Research Ethics Committee of the Faculty of Movement and Behavioural Sciences of the Vrije Universiteit Amsterdam (reference number: VCWE-2016-014).

Before study onset, participants were randomly allocated to either a text + personalized visual feedback or a text-only feedback condition, to control for individual differences in initial motivation and other potentially relevant background variables. A known issue with prolonged interaction with virtual agents is that it may become repetitive, leading to a decline in motivation and willingness to interact with a system (Bickmore et al., 2009b). Because our agent's design is fairly simplistic we wanted to control for this effect, and therefore, two weeks into the study, participants were randomized again to control for feedback history. Because no changes to the application could be made while the study was ongoing, randomization for the entire study took place before study onset. Participants were assigned to one of four possible groups, each with a different combination of text + personalized visual feedback (F) or text-only feedback (N) during weeks 1–2 and week 3.

2.2 Procedure

Before the study started, participants received an email with an invitation to fill out a digital informed consent form and demographic questionnaire. After giving their consent, participants installed the Android-only EMA application on their mobile devices. The application automatically stopped sending EMA requests once the study was over, after which it could be removed from the participants' smartphones.

2.3 Participants

As part of a research project for a bachelor's degree, students were asked to recruit at least 10 adults from their social network. Inclusion criteria for participation were (1) age 18 years or older, (2) owning an Android smartphone (minimal Android 2.3), and (3) no known severe mental health problems. Participants did not receive financial compensation, and were told that their participation would benefit the education of the student who had approached them.

2.4 Materials

2.4.1 Ecological Momentary Assessment of Mood

To collect self-monitored mood data, we built an Android smartphone application using the movisensXS EMA framework (movisensXS, Version 0.7.4162, Karlsruhe, Germany). The app prompted participants to rate their mood on their smartphone at three set time points each day (11:00, 15:00, and 20:00). Mood was assessed through the circumplex model of affect (Russell, 1980), which conceptualizes affective states as two-dimensional constructs comprising different levels of valence and arousal. Previous studies measured valence and arousal through 5-point scales (Asselbergs et al., 2016). We decided to tap both dimensions on a 3-point scale scored from –1 to 1 (negative to positive; low to high) (Figure 1), as this allowed a direct mapping to the visual feedback presented in the next section (Figure 2).

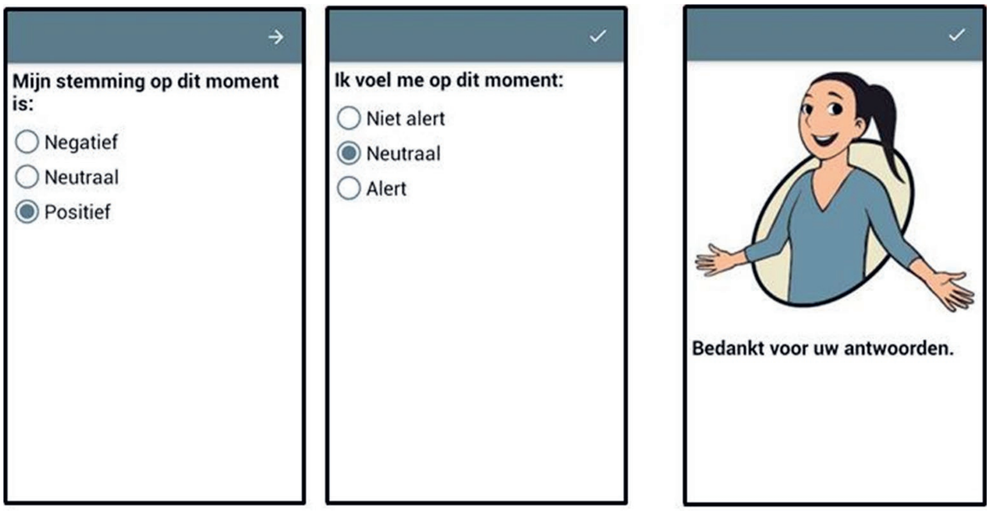

Figure 1. Screenshots of an EMA response and the system's reply including personalized visual feedback (translated from Dutch): left: "My mood at this moment is [Negative; Neutral; Positive]", middle: "At this moment I feel [Not alert; Neutral; Alert]", right: "Thank you for your answers

2.4.2 Personalized Visual Feedback with a Virtual Agent

After responding, participants received a message on a third screen, thanking them for their answers. In the personalized visual feedback condition, an embodied virtual agent accompanied this message. It consisted of the female version of the Pick-A-Mood (PAM) model (Desmet et al., 2012) that matched the reported mood. For example, a participant reporting positive valence (+1), and high arousal (+1), was deduced to be in an excited mood (Figure 2). Our visual feedback can be considered a simple form of empathy, where the system deduces and reflects users' moods based on their response to the EMA requests. With this feedback, we hoped to operationalize two concepts of Dialogue Support from Persuasive System Design (Oinas-Kukkonen and Harjumaa, 2009), the inclusion of which has been found to increase adherence (Kelders et al., 2012): social role ("if a system adopts a social role, users will more likely use it for persuasive purposes") by accompanying the thank you message with the face of a virtual character, and similarity ("people are more readily persuaded through systems that remind them of themselves in some meaningful way") by having the character mirror the participant-reported mood.

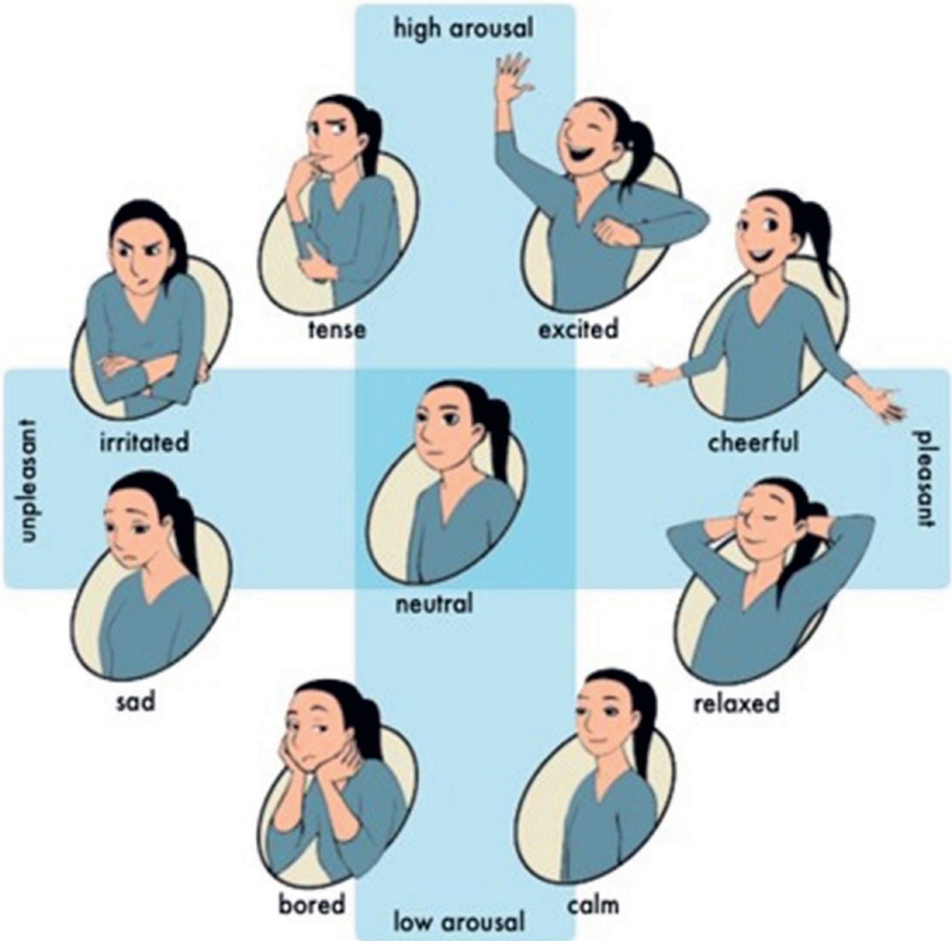

Figure 2. The Pick-A-Mood model, including an interpretation of mood ratings [Valence, Arousal]. Note that the model had to be tilted slightly clockwise to allow for the interpretation.

2.5 Measures

2.5.1 Adherence

Adherence was represented as a vector of binary values indicating either a response (=1) or no response (=0) to EMA requests on subsequent trials.

2.5.2 Feedback

Feedback was represented as a vector of binary values indicating either text + personalized visual feedback (=1), or text-only feedback (=0), accompanying the 'thank you for your answers' message.

2.5.3 Time

All trials received a sequentially ordered 'time stamp' ranging from the first to the last trial (Range = 1–63). Trials that were missing due to technical issues were added to the dataset, but with NA values for adherence, such that all 63 trials were accounted for.

2.6 Statistical Analysis

In our model, we assumed effects of time (people naturally lose interest in EMA after a while (Broderick et al., 2003)), feedback, and individual differences (some people may be more adherent to start with). To analyze the relationship between time and feedback as independent variables, and adherence as dependent variable, we used the *glmer* function from the *lme4* statistical package (Bates et al., 2015) in the R-environment (version 3.3.2) (R Core Team, 2016). For our main hypothesis, we conducted a logistic mixed effects analysis, with a 'feedback (1/0) x time (1–63)' interaction as fixed effect, and adherence (1/0) as the dependent variable. To investigate complex patterns over time, contrasts for the time variable were set to polynomials up to the tenth power. We accounted for individual differences at onset by adding random intercepts for the different participant IDs to our model. The regression model looked as follows in R-syntax:

$$\text{adherence} \sim \text{feedback} * \text{time} + (1|\text{id})$$

3 Results

3.1 Participant Flow

A total of 85 participants were recruited and randomized to one of the four groups, with 17 dropping out entirely or failing to install the software on time. From the 68 participants who had started, another 7 dropped out, and 7 were excluded from the analysis as they had experienced technical issues that resulted in either too few EMA requests (e.g., 0 to 2 per day on many different days), or too many (e.g., 4 per day) having been logged in the movisensXS data export files. Our final dataset included 54 participants (see Figure 3).

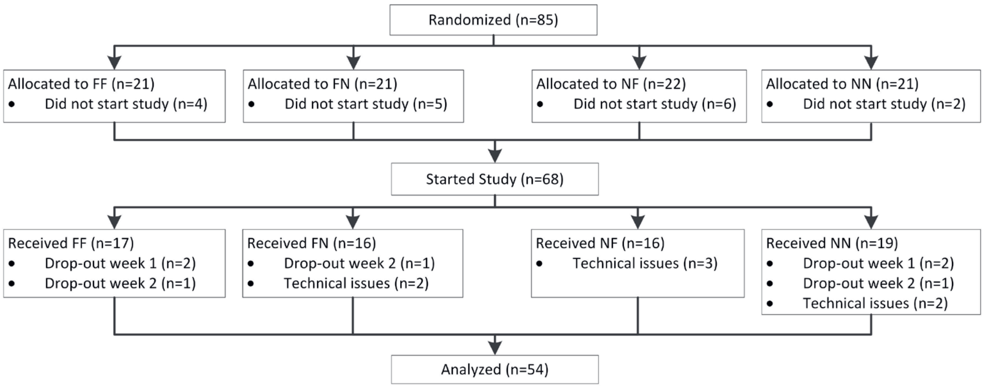

Figure 3. Participant flow

3.2 Descriptive Statistics

3.2.1 Participants

49 participants, 24 males and 25 females with mean age of 29.31 (Range = 20–64) filled out the demographic questionnaire, while 5 participants failed to do so.

3.2.2 Adherence

The 54 participants in the final dataset responded to EMA requests in 2004 out of a total of 3080 trials ($M = 37.1$ responses, and $M = 57.0$ trials per participant), resulting in an overall adherence of 65.1%. Note that the total number of trials does not add up to 63 on average (3 measurements per day for 21 days), as some requests were missed, either on the first day (e.g., when a participant started with the 20:00 measurement), or due to technical issues that prevented trials from taking place.

3.2.3 Feedback

The personalized visual feedback group received 1497 (48.6%) requests, and responded to 971 (64.9%). The text-only group received 1583 (51.3%) requests, and responded to 1033 (65.3%). The valence question was answered slightly more positively in the text-only condition, but no large differences existed between the two groups. With respect to the visual feedback provided, excited (n = 317 (32.7%)) and neutral (n = 241, (24.8%)) were most prominent.

3.2.4 Time

A total of 322 trials were unaccounted for due to technical issues, 183 of which would have contained personalized visual feedback. For our analysis, these trials were added to our dataset with missing values for adherence, giving us a total of 3402 trials. With the ten polynomial contrasts for the time variable our model was able to run, while visual inspection showed that no more than ten polynomials were to be expected.

3.3 The Effect of Feedback

3.3.1 Mixed Effects Logistic Analysis

Summary results of the mixed effects logistic analysis, based on all 3402 trials, are depicted in Table 1 below. Significant effects were found for feedback (p = .03), as well as the *feedback * time* interaction for the 3rd (p = .01), 7th (p = .01), and 9th (p < .01) order.

Table 1. Summary of the logistic mixed effects analysis results

Variable	Estimate	Standard Error	z-value	P-value
feedback	0.13	0.06	2.19	0.03*
time	0.09	0.35	0.27	0.79
feedback*time	0.07	0.42	0.17	0.86
feedback*time^2	0.56	0.37	1.49	0.14
feedback*time^3	0.93	0.37	2.53	0.01*
feedback*time^4	0.50	0.37	1.35	0.18
feedback*time^5	-0.42	0.36	-1.18	0.24
feedback*time^6	0.42	0.35	1.18	0.24
feedback*time^7	0.91	0.35	2.60	0.01*
feedback*time^8	-0.08	0.35	-0.23	0.82
feedback*time^9	-1.21	0.35	-3.45	0.00**
feedback*time^{10}	-0.01	0.35	-0.02	0.98

* $p < .05$, ** $p < .01$

3.3.2 Feedback * Time Interaction

Because of the significant interaction effects, the main effect of feedback cannot be interpreted as such. To further disentangle the model, the analysis was conducted again for both the feedback and no feedback condition, consequently leaving the feedback variable out of the equation. The results of these analyses with regard to the previously significant interaction effects are depicted in Table 2 below.

Table 2. Disentanglement of the significant interaction effects from the main analysis

	Variable	Estimate	Standard Error	z-value	P-value
visual feedback	$time^3$	-0.65	0.52	-1.26	.21
	$time^7$	-0.74	0.51	-1.45	.15
	$time^9$	0.69	0.50	1.38	.17
text-only feedback	$time^3$	1.05	0.52	2.01	.04*
	$time^7$	1.13	0.49	2.31	.02*
	$time^9$	-1.77	0.50	-3.55	.00**

* $p < .05$, ** $p < .01$

Whereas no significant effects of time were found for the feedback condition, the significant 3rd ($p = .04$), 7th ($p = .02$), and 9th ($p < .01$) order effects of time remained in the text-only feedback condition. A Chi-square test of the entire model further confirmed the significant *feedback * time* interaction effect (X^2 (62) = 84.99, $p = .03$). To visualize this, Figure 4 illustrates fluctuations in the no feedback condition (top left), which can be disseminated into separate polynomials of the 3rd (two bends; top right), 7th (6 bends; bottom left), and 9th (8 bends; bottom right) power. Meanwhile, the feedback condition is stable across all trials.

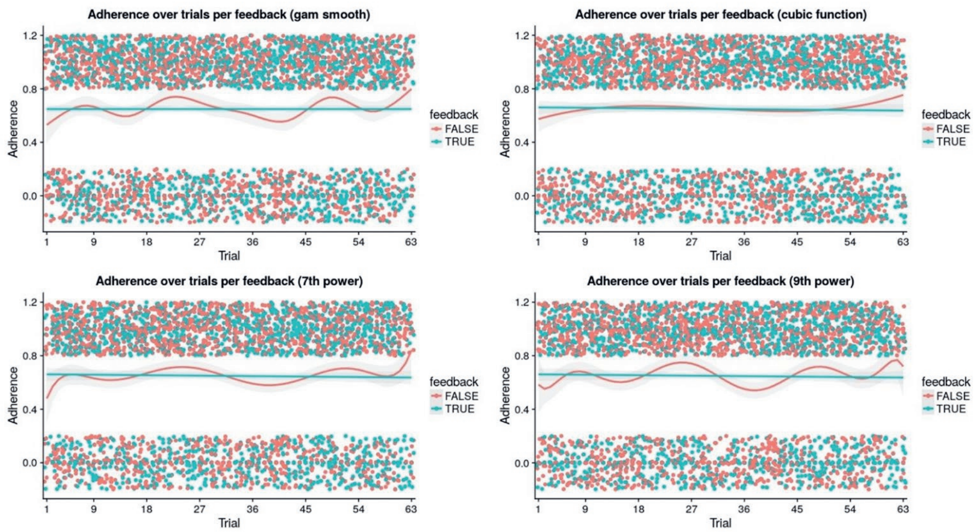

Figure 4. Visualizations of the overall model (top left), and the three significant polynomial effects, using a smoothed conditional means function

4 Discussion

4.1 Principal Results

Our visual feedback did not manage to increase EMA adherence, but rather kept it stable over time. A cautious interpretation of the results is that personalized visual feedback helped to maintain adherence to EMA in a more predictive manner than when text-only feedback was provided. Notably, as is visualized in Figure 4, there was no positive or negative linear trend for either condition. This is a surprising result given the assumption that adherence would decrease over time.

4.2 Implications

From a practical point of view, our results are interesting in the sense that from a researcher perspective, a constant flow of information, as provided by participants in the personalized visual feedback condition, may be preferable to the fluctuations observed in the text-only group. Mood, for example, changes relatively slowly over the course of more than one day. From a theoretical point of view, the results are harder to interpret, since the text-only group was at times more, and at times less adherent than the visual feedback group. Since contextual factors (e.g., having time for EMA responses) were controlled for by our randomization, the difference, i.e., fluctuation versus stabilization, can most likely be explained by an effect on participants' motivation. This would require further investigation; a possible future research paradigm to investigate the effect may include a qualitative component, e.g., by asking participants for their reasons to respond or not respond to EMA requests. It could also be interesting to look more closely at the moods that were reported per day as, for example, participants may be less inclined to answer requests when they are having a bad day compared to a good one.

Although personalized visual feedback seemed to have a stabilizing effect on adherence, it failed to increase it. It could be that personalized visual feedback does not matter that much, but it is also possible that the type of visual feedback we presented is not very effective, or cancelled out any positive effects. With regard to mood mirroring, for example, reflecting negative moods may actually amplify them (Pagliari et al., 2012), and thereby decrease user motivation. This was also argued following a study where a virtual agent mirrored users' emotional states to motivate them to play a game (Burleson, 2006), in which no significant effects on user motivation were found. Within the context of eMental Health interventions related to mood and our current paradigm, one interesting option for future research could be to investigate the effect of mirroring moods that have a negative valence with their equivalents on the positive side of the valence dimension. For example, a reported irritated mood (valence = –1, arousal = 1) would be mirrored by an excited mood (valence = 1, arousal = 1), which could alleviate the potential drawbacks of reflecting negative mood states.

4.3 Limitations

The virtual agent was a very simplistic one, which had to do with the technological limitations of the platform that we used to conduct the study. For example, the input for the 'reasoning engine' (Figure 2) of the embodied virtual agent was limited to the two questions that were answered in the current trial, and the EMA framework offered no options for more advanced animations. Although the agent was hardly impressive from a technological perspective, it did allow us to conduct research on its relationship to EMA adherence, within the limited degrees of freedom for agent design offered by a typical eMental Health framework. Additionally, we did so in a well-controlled paradigm: intervention with agent vs. intervention without agent.

It could also be questioned whether our specific feedback is the optimal one in this setting. We chose this feedback as it operationalized elements from Persuasive System Design theory (social role and similarity) known to be beneficial to adherence, and for pragmatic reasons related to the technological limitations of the EMA framework. Although it can be argued that similarity often refers more to user characteristics such as age or gender, we considered mood reflection relevant in our case since the task at hand was mood reporting. Some types of feedback that could be equally, if not more, relevant include reminders before EMA requests, targeting feedback only at participants who have been non-adherent for a period of time, or feedback that is specifically designed to uplift participants who report a negative mood.

A last limitation refers to the generalizability of our results. First and foremost, our study was conducted with a convenience sample, whose primary motivation for participation was likely to help out the students by whom they were recruited. Known mental health issues being an exclusion criterion, our study population was most likely quite different from the clinical populations that would typically use eMental Health interventions, and that may instead be motivated by a desire to improve their current situation. Additionally, there is still some debate as to whether EMA can be considered an intervention in itself (van Ballegooijen et al., 2016), which means generalizations to our broader context should be made with caution. These limitations are a natural consequence of the exploratory nature of our pilot study, but the methods we used could be applied equally well to contexts with real interventions and patients.

5 Conclusion

The study described in this paper was the first in a series of experiments which we hope will contribute to the empirical validation of the clinical effectiveness of embodied virtual agents in an eMental Health context. We aimed to find out whether feedback, operationalized by an embodied virtual agent, could increase adherence to mood rating requests in a three-week smartphone-based EMA study. While we did not find a significant main effect of feedback on adherence, there was a significant *feedback * time* interaction effect, which became apparent in fluctuations in adherence for the text-only condition, compared to a very consistent pattern in the personalized visual feedback group. To our knowledge, this paper represents one of the first explorative studies that used an embodied virtual agent, in a rigorous randomized and controlled design, to study a clinically relevant outcome measure over a prolonged period of time. Given the explorative nature and the relatively small sample size of this study, the stabilizing effect the virtual agent had on adherence has to be interpreted with some caution. Future studies may include a more sophisticated virtual agent, different feedback, a clinical study population, and a context more resembling CBT interventions.

3

CHAPTER 4

4

Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts: An Exploratory Study

Simon Provoost, Jeroen Ruwaard, Ward van Breda, Heleen Riper & Tibor Bosse (2019)

Frontiers in Psychology, 10:1065
doi: 10.3389/fpsyg.2019.01065

Abstract

Introduction: Sentiment analysis may be a useful technique to derive a user's emotional state from free text input, allowing for more empathic automated feedback in online cognitive behavioral therapy (iCBT) interventions for psychological disorders such as depression. As guided iCBT is considered more effective than unguided iCBT, such automated feedback may help close the gap between the two. The accuracy of automated sentiment analysis is domain dependent, and it is unclear how well the technology is applicable to iCBT. This paper presents an empirical study in which automated sentiment analysis by an algorithm for the Dutch language is validated against human judgment.

Methods: A total of 493 iCBT user texts were evaluated on overall sentiment and the presence of five specific emotions by an algorithm, and by 52 psychology students who evaluated 75 randomly selected texts each, providing about eight human evaluations per text. Inter-rater agreement (IRR) between algorithm and humans, and humans among each other, was analyzed by calculating the intra-class correlation under a numerical interpretation of the data, and Cohen's kappa, and Krippendorff's alpha under a categorical interpretation.

Results: All analyses indicated moderate agreement between the algorithm and average human judgment with respect to evaluating overall sentiment, and low agreement for the specific emotions. Somewhat surprisingly, the same was the case for the IRR among human judges, which means that the algorithm performed about as well as a randomly selected human judge. Thus, considering average human judgment as a benchmark for the applicability of automated sentiment analysis, the technique can be considered for practical application.

Discussion/Conclusion: The low human-human agreement on the presence of emotions may be due to the nature of the texts, it may simply be difficult for humans to agree on the presence of the selected emotions, or perhaps trained therapists would have reached more consensus. Future research may focus on validating the algorithm against a more solid benchmark, on applying the algorithm in an application in which empathic feedback is provided, for example, by an embodied conversational agent, or on improving the algorithm for the iCBT domain with a bottom-up machine learning approach.

1 Introduction

Internet-delivered cognitive behavioral therapy (iCBT) has been found equally effective as face-to-face therapy (Andersson and Cuijpers, 2009). It can be unguided or guided, with guidance being provided by trained volunteers or health professionals, and taking the form of “coaching”, e.g., providing motivation or technical assistance, or “treatment”, e.g., engaging in a therapeutic relationship (Schueller et al., 2017). Thus far, iCBT seems more effective when it includes guidance than when guidance is absent (Richards and Richardson, 2012; Andersson and Titov, 2014). The present study is part of a project in which we explore whether we can bridge the gap between guided and unguided interventions, either completely or partially, through automated support by embodied conversational agents (ECAs), computer-generated characters that can simulate verbal and non-verbal behaviors similar to those used in human face-to-face conversations (Isbister and Doyle, 2004).

When providing support, a number of non-specific factors are considered important. Examples are a good therapeutic alliance, positive expectancy effects of both the patient and the coach, therapeutic competence of the supportive human, and the content of written feedback (Mohr et al., 2011; Schueller et al., 2017; Mol et al., 2018). An important element in many of these factors is “empathy”, i.e., perceiving and understanding others’ affective states and acting accordingly (Paiva et al., 2017), which can contribute to a good therapeutic relationship between a supportive human and patient, both in face-to-face (Keijsers et al., 2000), and online settings (Mohr et al., 2011). Similar to human-human interaction, empathy simulated by an ECA can contribute to the bond between the user and ECA, as research has shown, for example, that empathic ECAs are seen as more trustworthy, likeable, and caring (Brave et al., 2005), and can build and sustain long-term relationships with users (Bickmore and Picard, 2005). For example, a recent study showed how the inclusion of empathy conveyed by a virtual character in a brief intervention for problematic drinking behavior could increase intentions to reduce drinking compared to the intervention without empathy (Ellis et al., 2017). In order to successfully express empathy, ECAs must accurately determine a user’s emotional state, and respond appropriately (Paiva et al., 2017). ECAs have been endowed with a variety of techniques to recognize the emotional state of the users they interact with. Examples are the analysis of facial expressions, body posture, acoustic features of speech, and other types of higher-level non-verbal behavior such as fidgeting, as well as linguistic content. In state-of-the-art ECA approaches, these features are combined to gain an optimal understanding of the user’s emotional state (Hartholt et al., 2013).

Internet interventions, however, are often website-based with limited possibilities for multi-modal interaction through audio-visual communication. Especially in unguided interventions, interaction is usually limited to patient responses to exercises or questions with multiple-choice (e.g., “Select an activity you will try to do this week”), Likert-scale (e.g., “Rate your mood on a scale of one to ten”), or free text input (e.g., “Please describe how you tried to apply the plan you made last week”). Rather than engaging in dyadic dialogs, as in face-to-face therapy or in the interaction with ECAs that make extensive use of emotion detection (e.g., (Swartout et al., 2013; Devault et al., 2014)), people who provide guidance in iCBT typically do so by email after a patient has completed one of several intervention modules, and base their messages on the patient’s input in the intervention and guidelines for giving online feedback (Mol et al., 2018). Thus, if we want an ECA to provide guidance similar to humans, it should respond to the scarce human input that is available in a guided internet-based intervention, rather than to information obtained during real-time dyadic interactions, and preferably do so empathically. Input based on a limited set of answers (e.g., rating one’s mood on a scale of one to ten) lends itself well to such a task, but dealing with free text input is more difficult as automated processing of the semantic nature of a text is still far from accurate, and often domain specific. In this paper we focus on the application of a technique that tries to determine the semantic content of a text on a higher level of abstraction, namely sentiment analysis.

Sentiment analysis, often interchangeably used with “opinion mining”, e.g., in the domain of product reviews, aims to identify text that contains sentiment, identify what the sentiment is, and determine the overall polarity (negative or positive) of the text (Pang and Lee, 2008). A sentence such as “I liked the module,” for example, would have a positive valence, while “I did not like the module,” would have a negative one. Broadly speaking, sentiment analysis algorithms are either based on bottom-up machine learning approaches, where algorithms learn to recognize sentiment by looking at example texts that have already been classified, while iteratively adjusting parameter values such that the algorithm’s output matches the predetermined classification, or top-down lexicon-based approaches, where they use pre-specified dictionaries to identify sentiment words (Medhat et al., 2014). Sentiment analysis has been researched extensively in the context of social media, in contexts ranging from predictions in politics (Sobkowicz et al., 2012) to the detection of depression (Wang et al., 2013). It has also been used in the area of ECAs, either by analyzing parsed speech or direct free text input, for example, to detect a user’s negative emotional states (see (Clavel and Callejas, 2016) for a review). An example of an application in the domain of psychology is the detection of depression in micro-blog posts (Wang et al., 2013).

From a literature review on the use of ECAs in clinical psychology, we learned that even though there have been a number of studies involving ECAs in a supporting role in iCBT contexts, evidence on their effectiveness, and validation of the underlying techniques that are used for the clinical domain remains sparse (Provoost et al., 2017). Clinical psychology is an applied science, and therefore, before novel technologies can be applied in practice, they require a thorough validation and understanding. However, we still know little about the application of sentiment analysis in the context of providing automated guidance in iCBT for depression. Moreover, sentiment analysis can be highly domain specific, considering, for example, the different vocabulary and grammatical engines required to classify newspaper articles versus Twitter messages. Therefore, in this paper we investigate how an existing sentiment analysis algorithm, using a top-down lexicon-based approach, evaluates free text input provided by patients using an iCBT intervention for depression compared to human judges. We further describe the implications of our findings for the applicability of sentiment analysis in clinical practice, and how they can inform further research.

2 Materials and Methods

2.1 Design

We conducted an exploratory study in which texts, written by patients during online therapy, were evaluated on overall sentiment and emotional expressiveness. Evaluation was conducted (1) automatically by an algorithm, and (2) manually by a group of human judges. From a set of 493 patient texts, subsets of 75 texts were randomly assigned to each human judge, such that every text received a similar amount of evaluations. Approval for the study was obtained from the Research Ethics Committee of the Faculty of Movement and Behavioral Sciences of the Vrije Universiteit Amsterdam (Reference Number: VCWE-2017-165).

2.2 Participants

A total of N=52 first-year psychology students at Vrije Universiteit Amsterdam were recruited. As part of their curriculum they have to partake in experimental studies for at least 10 h, and they received 1 h worth of study participation credits as compensation.

2.3 Materials and Measures

2.3.1 Patient Texts

The texts we used were part of the patient input in a randomized controlled trial that compared a blended internet-based CBT intervention for major depression to an established face-to-face CBT treatment in specialized mental health care settings (Kooistra et al., 2014). Blended, in this case, refers to the integration of an online intervention with weekly face-to-face conversations with an assigned therapist. The online component consisted of 10 modules, typically containing a mixture of psychoeducation and therapeutic exercises. At the end of each module, except for the introductory one, patients were asked for a non-obligatory evaluation on how they experienced the internet module. Therapists were instructed to focus their online communication and feedback on the progress patients made within the intervention, rather than on what had occurred during the face-to-face conversations. The texts were anonymized, and of the texts that occurred more than once, only one was kept in the dataset.

2.3.2 Automated Sentiment Analysis

To analyze the texts' sentiment, we made use of a sentiment mining algorithm that has been tailored toward the Dutch language by using a Dutch grammatical engine and vocabulary (Sentimentics, Rotterdam, The Netherlands). Our choice for this domain-independent algorithm was a pragmatic one, since we are exploring a new domain where tailored algorithms have not yet been developed, and state-of-the-art algorithm's like those developed by Google (Google, 2018) or IBM (IBM, 2018) have, at the time of writing, not been tailored to the Dutch language. The algorithm has previously been used in studies focusing on crime prediction, where it was used to identify aggressive Twitter messages (Gerritsen and Van Breda, 2015), life-style support, where it was used to determine people's attitude toward a lifestyle goal (Van Breda et al., 2012), and a training application for football referees, where it was used to identify the language used by referees during conflicts with football players (Bosse et al., 2017). Though the systems that used the algorithm seemed to have potential following simulation studies and a preliminary evaluation, respectively, they are still under development, and none of the published studies specifically targeted the accuracy of the algorithm. The algorithm accepts strings of text as input, and returns, among others, an overall sentiment value, and scores for 33 specific emotions contained within the text. It uses an advanced form of the bag-of-words approach (Raghavan and Wong, 1986; Zhang et al., 2010) by utilizing lists of words with vectors of weights, and combines this with a rule-based system, operationalized through a grammar detection engine, to analyze the surrounding semantic context of the found sentiment words.

2.3.2.1 Overall sentiment

When positive or negative words that are present in the lexicon are identified in the text, the algorithm looks for words that represent the semantic context surrounding them with a grammatical engine. Such words include negation (e.g., "not"), strengthening (e.g., "more") and weakening (e.g., "less") words. To all sets of word sequences found a scoring method is applied. The vectors of weights for each word sequence are multiplied with a particular weight parameter, and fed to a tangent or logistic activation function. This function normalizes the value to one within the interval [–1, 1], and ensures that extreme scores do not have a disproportionately large effect on the final scores. In the final independent score for positive and negative sentiment, –1 means very negative, and 1 very positive. An example is provided in Figure 1.

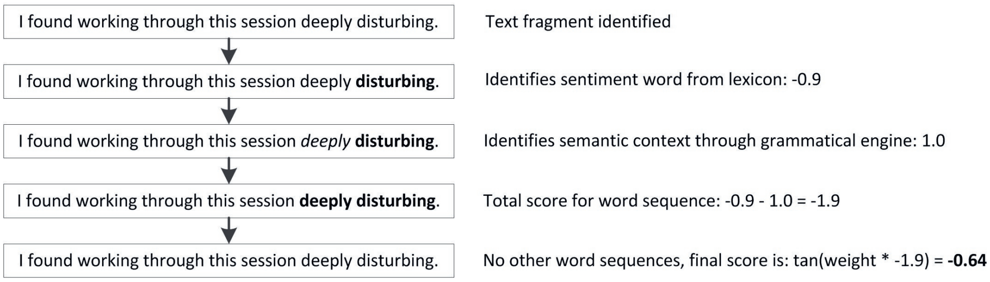

Figure 1. Sentence exemplifying how the algorithm comes to an overall sentiment score for a sentence containing one sentiment word ("disturbing") from the lexicon combined with a strengthening word ("deeply") identified by the grammatical engine

2.3.2.2 Emotional expressiveness

For the emotion labels, a similar approach is taken, be it that the text is now searched for words related to that emotion instead of sentiment words in general as a first step. Additionally, it considers emotions either not present or present to a certain extent, and therefore applies a different weight parameter to the emotion labels. Scores lay in the interval [0, 1], where 0 means the emotion was not detected, and 1 means the maximum amount of the emotion was detected.

2.3.3 Human Sentiment Analysis

An online questionnaire was designed for the human judges, in which they were presented with a different page for each patient text. For every text, they were asked to evaluate overall sentiment and the presence of five emotions with the use of a slider (see Figure 2).

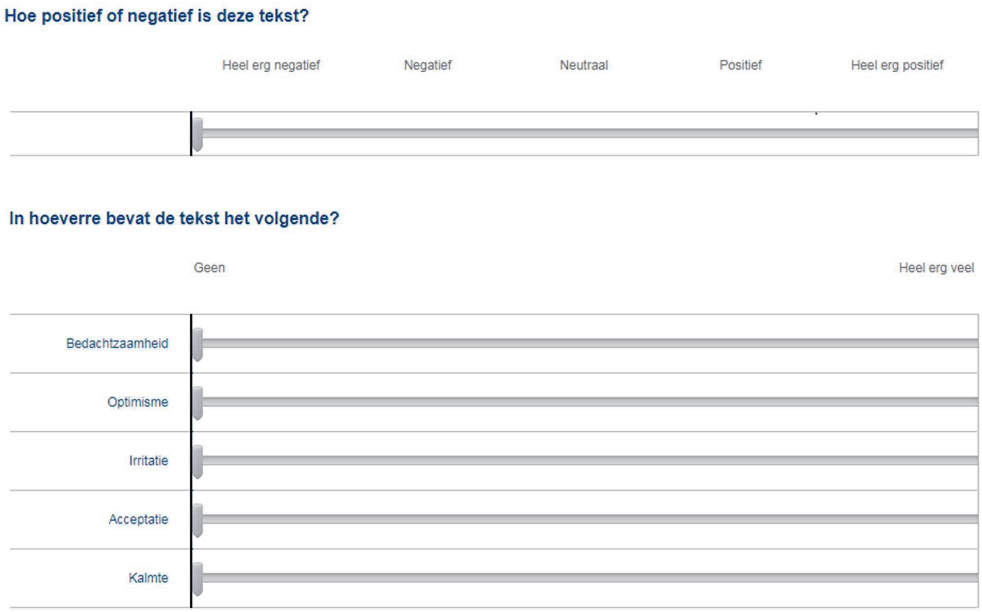

Figure 2. Sliders used by human judges to evaluate texts on sentiment and emotions. Translated from Dutch: (top) "How positive or negative is this text?", answer labels: very negative; negative; neutral; positive; very positive, and (bottom) "To what extent does the text contain the following?", answer labels: none; a whole lot, five emotions top to bottom: pensiveness; optimism; annoyance; acceptance; serenity

Because asking our participants to evaluate the texts on all 33 emotions considered by the algorithm would have been too burdensome, we chose to focus on the five emotions that the algorithm detected in the texts most often. The summed algorithm scores over all 493 texts are depicted in Figure 3, and the five most prominent emotions we chose to study were pensiveness, annoyance, acceptance, optimism, and serenity. To prepare the data for the ICC analysis, the ratings provided by human judges were scaled to the same intervals as the automated ratings, i.e., [−1, 1] for overall sentiment, and [0, 1] for the five emotions.

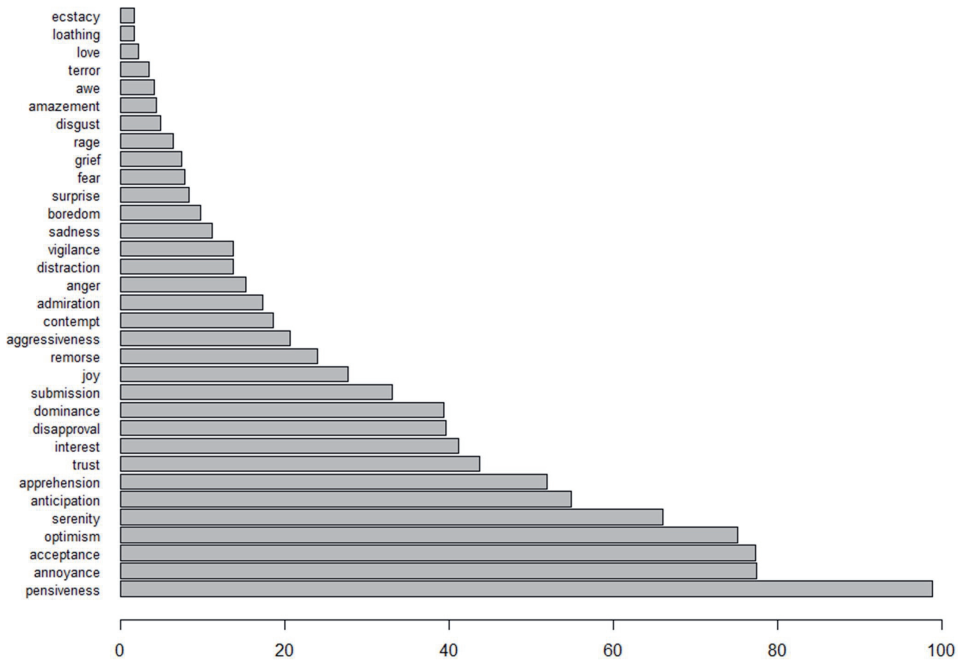

Figure 3. The summed algorithm scores over all 493 texts, with respect to all 33 detectable emotions

2.4 Procedure

Participants could register for the study during March and April 2018, and those that did received a personal link to the online questionnaire. At the beginning of the questionnaire they received instructions and signed a digital informed consent, and at the end were given a debriefing with more information about the study.

2.5 Analysis

To compare algorithm with human judgment, we first calculated the average human rating for every text, which was then used as the value for human judgment. We also converted the numerical values to categorical ones, as practical applications may use a limited set of observable user responses (e.g., positive or negative) to be matched by a limited set of agent responses, for example by responding with positive reinforcement to positive or empathically to negative user input. This is similar to the approach where polarity can be measured either by intensity (continuous) or direction (categorical) (Devitt and Ahmad, 2007). For overall sentiment, three categories were created: negative for values smaller than 1, positive for values larger than one, and neutral for values equal to zero. Because the comparison between humans and algorithm uses the human judges' average ratings, which are unlikely to be equal to zero, we defined neutral sentiment as any value equal to or larger than –0.1, and equal to or smaller than 0.1. For the different emotions, two categories were created: not present for values equal to or smaller than 0.1, and present for values larger than 0.1.

We used the R statistical software (R Core Team, 2016), extended with the *psych* (Revelle, 2018) and *irr* (Gamer et al., 2012) packages to calculate intra-class correlations (ICC), Cohen's kappa (κ), and Krippendorff's alpha (α).

2.5.1 Intra-class correlation

ICCs can be used to assess IRR on items with a continuous scale, and are also applicable to cases with more than two raters or missing values. Values range between 0 and 1, with values less than 0.5 indicative of poor, values between 0.5 and 0.75 of moderate, values between 0.75 and 0.9 of good, and values greater than 0.9 of excellent reliability (Shrout and Fleiss, 1979). In our analysis under the continuous interpretation of the results, we used ICC(1,1) to calculate agreement among humans, accounting for one-way random effects caused by our research setup (different judges from one population rate each item), and ICC(3,1) to calculate agreement between the algorithm and human judges, accounting for two-way mixed effects (Koo and Li, 2016).

2.5.2 Cohen's kappa

Cohen's kappa can be used to assess IRR on items with a categorical scale (Cohen, 1960), and a weighted κ can be calculated to account for ordered categories. Values range between –1 and 1, with values lower than 0 indicative of no, values between 0 and 0.4 of slight to fair, values between 0.41 and 0.6 of moderate, values between 0.61 and 0.8 of substantial, and values between 0.81 and 1 of almost perfect reliability (Landis and Koch, 1977). In our analysis we used the weighted κ to calculate agreement between the algorithm and humans under the categorical interpretation of results.

2.5.3 Krippendorff's alpha

Krippendorff's alpha can also be used to assess IRR on items with a categorical scale, but contrary to Cohen's Kappa is able to deal with missing values (Hayes and Krippendorff, 2007). Values range between –1 and 1, and can be considered reliable if larger than 0.8, with values larger than 0.67 allowing tentative conclusions to be drawn (Artstein and Poesio, 2008). We used α to calculate agreement among human judges under the categorical interpretation of results, since there were many missing values due to every item being rated by a limited set of judges.

3 Results

3.1 Descriptives

A total of 52 participants completed the experiment. Every text received on average $M = 8.1$ (range 4–10) human evaluations, and a total of 3900 evaluations was provided. Table 1 contains the summary statistics of the averaged human and algorithm evaluations with respect to sentiment and the five emotions.

Table 1. Summary statistics of averaged human ($M = 8.1$ evaluations per text) and algorithm judgment of the N=493 total number of texts

	M	Min.	Max.	*SD*
Sentiment				
Human Judges	0.00	-0.71	0.67	0.30
Algorithm	0.04	-0.97	0.99	0.44
Pensiveness				
Human Judges	0.47	0.01	0.78	0.15
Algorithm	0.47	0.00	1.00	0.44
Annoyance				
Human Judges	0.40	0.04	0.81	0.17
Algorithm	0.39	0.00	1.00	0.44
Optimism				
Human Judges	0.37	0.01	0.86	0.20
Algorithm	0.40	0.00	1.00	0.42
Acceptance				
Human Judges	0.25	0.02	0.86	0.17
Algorithm	0.40	0.00	1.00	0.42
Serenity				
Human Judges	0.38	0.02	0.79	0.16
Algorithm	0.37	0.00	1.00	0.41

It shows a smaller range of values for human evaluation, caused by taking the average values, and consequently a smaller standard deviation as well. For human evaluation, median values did not differ much from mean values, which means there was no reason to use the median in our analysis.

Figure 4 depicts the distributions of the ratings, with those for sentiment approaching a normal distribution, and those for the different emotions showing right-skewed distributions with one large peak (no emotion detected) for the algorithm, and varying distributions for human evaluation.

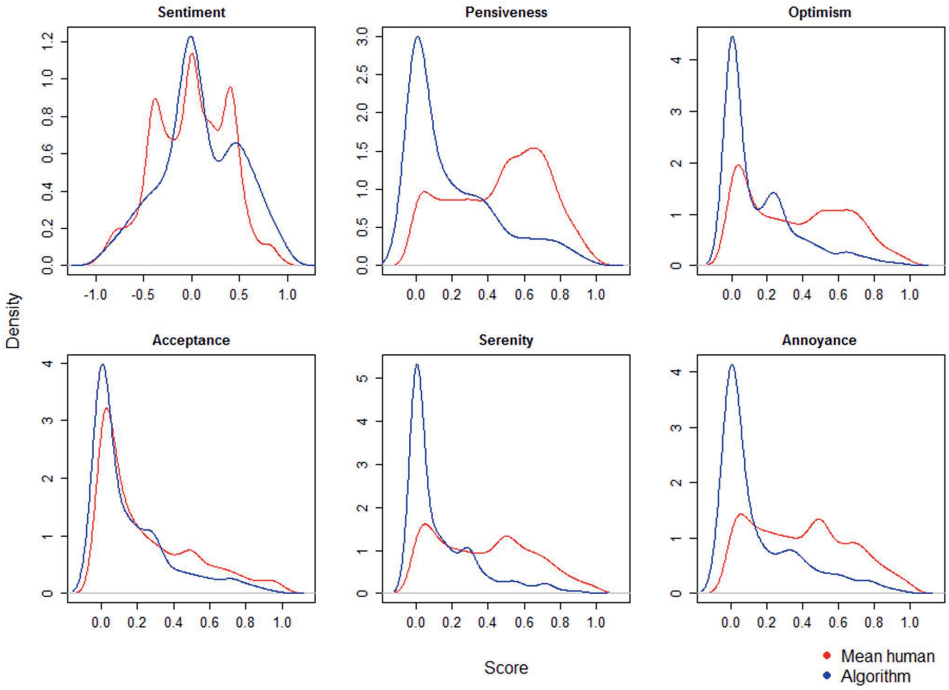

Figure 4. Probability distributions of human and algorithm evaluations

Figure 5 shows scatter plots of the raw data including a line representing a linear fit model. On visual inspection, human and algorithm evaluation have the best correlation for sentiment, while for acceptance and annoyance they are negatively correlated.

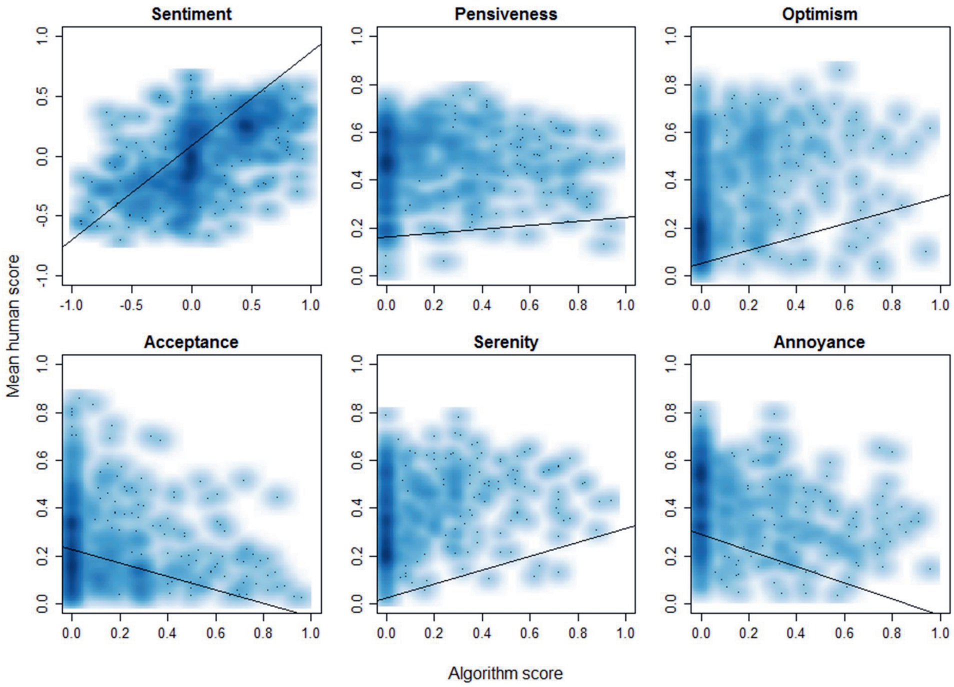

Figure 5. Scatterplots of the algorithm versus the mean human scores for all of the 493 patient texts, including a line representing the linear fit model.

4

3.2 Agreement Among Human Raters

3.2.1 Continuous Interpretation

As shown in Table 2, an ICC(1,1) analysis revealed a moderate IRR among human raters with regard to sentiment (ICC = 0.58, CI = 0.54–0.61, $F(492,25143) = 71$, $p < 0.01$). IRR among human raters with regard to the different emotions was overall poor, and only with regard to optimism did the judges get close to moderate agreement.

Table 2. Results of intra-class correlation on human-human agreement

	Intraclass Correlation	95% confidence interval		F-test with true value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
Sentiment	.58	.54	.61	71	492	25143	.00
Pensiveness	.22	.20	.25	16	492	25143	.00
Annoyance	.28	.16	.31	21	492	25143	.00
Optimism	.46	.43	.50	46	492	25143	.00
Acceptance	.34	.32	.38	28	492	25143	.00
Serenity	.24	.21	.26	17	492	25143	.00

3.2.2 Categorical Interpretation

For the categorical human-human interrater agreement, a Krippendorff's alpha of $\alpha = 0.51$ was found, which indicates moderate agreement similar to ICC results for the continuous interpretation of ratings. Agreement with respect to the emotions was poor. An overview is provided in Table 3.

Table 3. Krippendorff's alpha values for human-human agreement

Emotion	**α**
Sentiment	.51
Pensiveness	.17
Annoyance	.20
Optimism	.28
Acceptance	.23
Serenity	.17

3.3 Agreement Between Human Judges and Algorithm

3.3.1 Continuous Interpretation

An ICC(3,1) analysis revealed a moderate IRR between human raters and the algorithm with regard to sentiment (ICC = 0.55, CI = 0.48–0.61, $F(492,492) = 3.4$, $p < 0.01$). IRR between the human raters and the algorithm with regard to the different emotions was poor (see Table 4).

Table 4. ICC values for human-algorithm agreement.

	Intraclass Correlation	95% confidence interval		F-test with true value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
Sentiment	.55	.48	.61	3.4	492	492	.00
Pensiveness	.12	.03	.21	1.3	492	492	.00
Annoyance	.00	-.09	.09	1	492	492	.5
Optimism	.23	.14	.31	1.6	492	492	.00
Acceptance	.00	-.09	.09	1	492	492	.5
Serenity	.14	.06	.23	1.3	492	492	.00

3.3.2 Categorical Interpretation

A weighted Cohen's kappa was calculated, indicating there was moderate agreement with regard to sentiment, $\kappa = 0.58$ (95% CI, 0.52 to 0.63), $p = 0.05$. As is illustrated in Table 5, similar to the agreement observed when considering continuous data, the correlation between human and algorithm evaluation remained poor.

Table 5. Cohen's kappa for human-algorithm agreement regarding emotions

Emotion	κ	95% CI
Sentiment	.58	.52 to .63
Pensiveness	.01	-.01 to .03
Annoyance	-.01	-.04 to .02
Optimism	.09	.03 to .14
Acceptance	-.07	-.14 to .00
Serenity	.03	.00 to .06

Tables 6 and 7 give an overview of the accuracy of the algorithm compared to human judgment for both overall sentiment (65.92%) and the emotion with the highest κ-value, optimism (54.8%).

Table 6. Comparison of the average human judges' and algorithm's evaluations with respect to sentiment

		Algorithm		
		negative	neutral	positive
Human Average	negative	144	15	38
	neutral	30	19	41
	positive	25	19	162

Table 7. Comparison of the average human judges' and algorithm's evaluations with respect to optimism

		Algorithm	
		not present	present
Human Average	not present	32	12
	present	211	238

4 Discussion

4.1 Principal Results

The aim of this study was to investigate how well humans and a sentiment analysis algorithm agree on evaluating the overall sentiment and presence of five emotions in patient input in an iCBT intervention for depression. Regarding sentiment, human-human agreement was moderate, both under the continuous (ICC = 0.58) and categorical (α = 0.51) interpretations. Algorithm-human agreement was moderate as well, again both under the continuous (ICC = 0.55) and categorical (κ = 0.58) interpretations. With respect to the different emotions, human-human agreement was overall poor. Most consensus was achieved on optimism (ICC = 0.46), which could be considered 'fair' agreement under an alternative interpretation (Cicchetti, 1994). Human-algorithm agreement on the presence of emotions was poor as well, with the highest agreement once again being achieved for optimism (ICC = 0.23), and agreement on the other emotions being negligible. In two cases, for acceptance and annoyance, the correlations even appeared to be slightly negative, as can be observed in Figure 5. The interpretation of low human-algorithm agreement on the presence of emotions deserves some caution though, as the low human-human agreement for the presence of emotions compared to the moderate human-human agreement for overall sentiment makes for a less solid benchmark.

When we look at performance of the algorithm compared to human judgment under the categorical interpretation in terms of accuracy (65.92%) it seemed reasonable, but higher values have been reported. Some examples of higher accuracies are 70.2% for a corpus of software reviews when comparing an algorithm to three judges (Aggarwal and Singh, 2013), and around 88% for a corpus of 1000 comments on Youtube videos related to anorexia (Aggarwal and Singh, 2013) and some with lower accuracies are 47% for a corpus of news reports on a potential hostile take-over of an airline (Devitt and Ahmad, 2007), and 65.7% for 447 subjective statements in a corpus of 10 news articles (Wilson et al., 2005). Algorithms for the Dutch language have been benchmarked sparsely, an example being 70.0% for a corpus of 60 social media texts (Tromp, 2011). Since there were an average of M = 8.1 human ratings per text, and the mean human rating for each text was used as a benchmark, calculating an accuracy score for human-human agreement is not straightforward. Nevertheless, we can say that there was relatively low agreement among human judges, as Wilson et al. in their study using 10 news articles, for example, reported an accuracy of 82%, and IRR of κ = 0.72 with regard to evaluating sentiment (Wilson et al., 2005).

4

With respect to overall sentiment, this means that the algorithm does about as well at discerning between positive and negative sentiment as would a random human judge from our population of judges, agreement being moderate in both conditions. The same thing held for the evaluation of emotions, as agreement was poor in both conditions. This could mean that it is equally difficult for humans to evaluate the texts in our domain, or that the emotions we chose are hard to apply to our texts, but it could also indicate that the texts are too ambiguous. Even though they were supposed to be about patients' opinions on the treatment module they had just finished, they were not necessarily limited to this topic. Since we used texts from a blended-care intervention, i.e., patients spoke to their therapists face-to-face as well, it is possible that patients took the review as an opportunity to inform their therapists of other things as well, such as significant life events that took place. On the other hand, however, it is also possible that it was the other way around, with patients preserving emotionally laden topics for the face-to-face meetings. Both factors may have been at play, as texts varied a lot in terms of length and content, with input ranging from the Dutch abbreviation for 'not applicable' and concise evaluations of the module (e.g., "This session gave me a good insight into different activities"), to extensive reports on their current mental state and recent life events, such as the impact of family events on their mood. Despite the ambiguous nature of some texts, we chose to use them as they were, because manual filtering of the texts would not happen in the practical application we envision either.

Furthermore, we chose to benchmark the algorithm versus the average scores of human judges as a "gold standard." These human judges, however, were first year psychology students, and although these can be considered to have at least some affinity with our domain, actual therapists or people with experience in providing guidance to iCBT as human judges may have provided data more closely resembling the evaluation of judgment in clinical practice. However, that this would not necessarily mean higher inter-rater agreement, is exemplified by two studies into inter-rater agreement among clinicians when identifying overt problems and underlying schemas of CBT patients (Persons et al., 1995; Persons, Jacqueline B.; Bertagnolli, 1999). Even though the judges in these studies were asked to rate patient interviews instead of short text fragments, the studies showed that clinicians can have considerable trouble agreeing on patient data as well. This was especially true when individual judges were concerned instead of a group average, and little evidence was found that a higher level of expertise improved judgment.

4.2 Future Research

With average human judgment as our "gold standard," and the algorithm performing about as well as a random human judge from our population, the algorithm could be applied for overall sentiment analysis of patient input in a practical application if we consider this "gold standard" as good enough. This is not the case for the specific emotions, since both human-human and human-algorithm agreement were low at best. As briefly described in the introduction, our aim is to develop an ECA that can support people who are working through an iCBT intervention for depression. Because of the technological limitations in interpreting and producing natural language automatically, we envision a tree-based dialog approach in which users choose their responses from a menu, as the safest way to structure a conversation. The limited number of pathways that represent all different possible conversations are relatively easy to understand and finite, which is important if we want mental health specialists to review the dialogs. Moreover, tree-based dialog approaches allow for the use of threshold values of parameters (e.g., negative or positive sentiment) to determine which path to take through a conversation, and therefore lend themselves well to our domain.

Considering such a supportive ECA, applying the algorithm to descriptions of how patients experienced intervention modules, could allow it to determine whether this was positive or negative, and to consequently provide automated personalized feedback through positive reinforcement (e.g., "It seemed to me that you liked working through the previous module, good luck with the next one!") or a more empathic response (e.g., "I noticed you finished the previous module, even though you did not seem to like it much. Impressive, keep up the good work!") respectively. Given the relatively large margin of error (around 1 in 3 evaluations would be wrong given our 65.92% accuracy), however, it seems imperative to build in a security mechanism to avoid incoherent communication by the ECA. A possibility is to first ask a user for confirmation (e.g., “It seemed to me that you did not like the previous module. Is that correct?”), and only then continue with the appropriate motivational message. Incorporating such a mechanism in a dialog may seem superfluous, as the ECA could also ask a patient directly about his or her opinion on the previous module, however, an important aspect of providing support is for a guiding person to show that he or she has actively looked at what a patient has been doing, for example through summarizing or reflecting (Mol et al., 2018), rather than asking things that a patient has already told the guiding person indirectly through the intervention.

We could also look to apply the algorithm to, or validate it against other user input, such as responses to questions about willingness and readiness to change, or to questions about how their homework exercises (e.g., experimenting with certain behaviors) are going. By targeting the right set of patient input, it may even be possible to determine a general mood of the patient by comparing the use of language over time.

A different angle of future work would be to tune the algorithm to our iCBT domain. Now that we have a large dataset of human evaluations, a bottom-up approach with machine learning could be used to tailor the algorithm toward the domain-specific vocabulary. Although to our knowledge, this has not been done in the context of sentiment analysis in iCBT, machine learning has been successfully applied in the domain of mental health disorders. For example, data-driven approaches outperformed human judgment in the prediction of iCBT treatment success (Amethier and Haggren, 2018), and non-fatal suicide attempts (Walsh et al., 2018). A comparison with these approaches deserves some caution, however, firstly because the data which was used to train the algorithms consisted of demographic and psychometric information rather than text. Secondly, treatment success, as well as non-fatal suicide attempts, are more solid benchmarks than human judgment in this study. If we aim to outperform a gold standard, in our case one of judging sentiment contained in a text, a more objective benchmark may be required. In light of the sparse evidence regarding the validity of human judgment as a benchmark, it could be interesting to compare an algorithm’s performance to patients’ own judgment.

5 Conclusion

Sentiment analysis could be a promising tool with which to enhance the personalization of automated feedback in iCBT interventions, for example through conversations with a supportive ECA. Our study showed that an existing algorithm for the Dutch language performed about equally well as a randomly chosen human judge at distinguishing between negative, neutral, and positive sentiment present in free-text patient input. The algorithm performed poorly at evaluating the presence of specific emotions, but the human judges, even though they were more consistent with each other than with the algorithm, performed poorly as well in terms of inter-rater agreement. This means that it may be worthwhile to validate the algorithm against a potentially more solid benchmark, such as patients' own judgment. If we were to consider the level of human-human agreement reported in this study to be the gold standard for our domain, automated sentiment analysis could be considered applicable. However, given the somewhat higher accuracy scores found in the analysis of, for example, social media messages or product reviews, it may be worthwhile to build in a security mechanism that confirms the automated analysis if it were used in practice, or to tailor an algorithm to the domain of iCBT interventions.

Acknowledgments

The authors would like to express their sincere thanks to Ms. Lisa Kooistra for her assistance in acquiring the texts that were evaluated.

CHAPTER 5

5

Behind the Scenes of Online Therapeutic Feedback in Blended Therapy for Depression: Mixed-Methods Observational Study

Mayke Mol, Els Dozeman, Simon Provoost, Anneke van Schaik, Heleen Riper & Johannes Smit (2018)

Journal of Medical Internet Research 20(5) e174
doi: 10.2196/jmir.9890

Abstract

Background: In Internet-delivered cognitive behavioral therapies (iCBT), written feedback by therapists is a substantial part of therapy. However, it is not yet known how this feedback should be given best and which specific therapist behaviors and content are most beneficial for patients. General instructions for written feedback are available, but the uptake and effectiveness of these instructions in iCBT have not been studied yet.

Objective: This study aimed to identify therapist behaviors in written online communication with patients in blended CBT for adult depression in routine secondary mental health care, to identify the extent to which the therapists adhere to feedback instructions, and to explore whether therapist behaviors and adherence to feedback instructions are associated with patient outcome.

Methods: Adults receiving blended CBT (10 online sessions in combination with 5 face-to-face sessions) for depression in routine mental health care were recruited in the context of the European implementation project MasterMind. A qualitative content analysis was used to identify therapist behaviors in online written feedback messages, and a checklist for the feedback instruction adherence of the therapists was developed. Correlations were explored between the therapist behaviors, therapist instruction adherence, and patient outcomes (number of completed online sessions and symptom change scores).

Results: A total of 45 patients (73%, 33/45 female, mean age 35.9 years) received 219 feedback messages given by 19 therapists (84%, 16/19 female). The most frequently used therapist behaviors were informing, encouraging, and affirming. However, these were not related to patient outcomes. Although infrequently used, confronting was positively correlated with session completion ($\rho = .342$, $p = .02$). Therapists adhered to most of the feedback instructions. Only 2 feedback aspects were correlated with session completion: the more therapists adhere to instructions containing structure (limiting to 2 subjects and sending feedback within 3 working days) and readability (short sentences and short paragraphs), the less online sessions were completed ($\rho = -.340$, $p = .02$ and $\rho = -.361$, $p = .02$, respectively). No associations were found with depression symptom change scores.

Conclusions: The therapist behaviors found in this study are comparable to previous research. The findings suggest that online feedback instructions for therapists provide sufficient guidance to communicate in a supportive and positive manner with patients. However, the instructions might be improved by adding more therapeutic techniques besides the focus on style and form.

1 Introduction

1.1 Internet-Delivered Cognitive Behavioral Therapy

There is considerable evidence that Internet-based interventions are effective for the treatment of mild, moderate, and major depression (Andersson and Cuijpers, 2009; Andrews et al., 2010; Josephine et al., 2017). Therapist-guided, Internet-delivered cognitive behavioral therapy (iCBT) has been found to be more effective than unguided iCBT (Andersson and Cuijpers, 2009; Richards and Richardson, 2012) and has also been found to be equally effective compared with face-to-face-delivered CBT (Carlbring et al., 2018). Beyond these findings, a number of studies focused on nonspecific factors that might be effective in iCBT (e.g., therapeutic alliance, therapist competence, and placebo-expectancy effects) and especially showed interest in the role of therapist guidance in iCBT (Schueller et al., 2017). So far, mixed results have been found. In a systematic review, Richards and Richardson, e.g., found that the way guidance is given has an impact on treatment adherence in depressed patients (Richards and Richardson, 2012). Therapist-guided iCBT had a 72% completion rate, iCBT interventions with administrative support (support by staff to guide patients through the program in a nontherapeutic way) 65%, and interventions with no support at all 26%. For Internet-delivered problem-solving treatment (PST), there also is evidence that the level of support is important in reaching effects for patients with depression (Kleiboer et al., 2015). Patients who received PST with weekly support from a coach improved significantly more than the waitlist control group. In the group that received no support, completion rates were lowest (22%), and the completion rates were highest in the group that received nonspecific support (60%). Patients who received weekly support had comparable completion rates with patients who received "support on request" (33% and 31%, respectively). In a study by Titov et al, patients with depression showed significant clinical improvement after receiving iCBT, regardless of whether the support came from a therapist or a technician (Titov et al., 2010). Findings of a recent American study indicate that iCBT with 5 hours of therapeutic face-to-face contact was noninferior to CBT that provided over 8 additional hours of therapist contact for patients with depression (Thase et al., 2018).

5

1.2 Therapist Behaviors

However, there is much more to discover about online guidance. One point of interest is how therapists give online feedback to their patients. This can be done by looking at the communication strategies and content they use in their written support. For example, looking at therapist behaviors such as validating what patients write (e.g., "That must be very difficult for you...") and stimulating patients to come up with their own solution (e.g., "When was the last time you felt that way? What did you think and what did you do differently?"). Written feedback is a substantial part of Internet-based treatments and requires specific skills of therapists. It is therefore interesting to further explore such therapeutic microprocesses in online feedback because this part of therapy may be very relevant in the adherence and also the effectiveness of iCBT (Johansson and Andersson, 2012).

1.3 Therapist Behaviors in Face-to-Face Therapy

The content of feedback and its impact on treatment results have been studied in face-to-face–delivered psychotherapies, and especially in CBT. Studies have identified different therapist behaviors that are frequently used in CBT sessions with patients. These behaviors range from expressing empathy, making supportive communications (e.g., encourage, praise, or guide the patient), asking directive questions, and confronting patients with different points of view (Keijsers et al., 2000; Watson and McMullen, 2005; Watzke et al., 2008). Self-disclosures by therapists appear to be infrequently used (Keijsers et al., 2000), although these are generally considered helpful by patients in the therapeutic process (Hill et al., 1988; Knox and Hill, 2003). In addition, research shows that therapist behaviors such as expressing empathy, giving positive regards, and confronting patients can have a positive impact on treatment outcome in CBT for various patients such as people with depression (Keijsers et al., 2000).

1.4 Therapist Behaviors in Internet-Delivered Cognitive Behavioral Therapy

Therapist behaviors in iCBT have also been studied. This was done for several psychiatric diseases such as eating disorders (Sánchez-Ortiz et al., 2011), insomnia (de Bruin and Meijer, 2017), anxiety (Paxling et al., 2013), and depression (Holländare et al., 2016; Schneider et al., 2016). Comparable to the behaviors in face-to-face therapies, the most frequently used therapist behaviors were encouraging, reinforcing, and supporting patients. When looking at the association between these therapist behaviors, patient treatment outcome, and patient online session completion, mixed results were found. Holländare et al found that encouraging, guiding, and affirming were strongly associated with session completion (Holländare et al., 2016). Encouraging, affirming, and self-disclosure were weakly to moderately associated with an improvement in depressive symptoms. The most important finding by Paxling et al was the effect of therapists' task reinforcement (e.g., reinforcing completed assignments) on session completion as well as treatment outcome (Paxling et al., 2013). Interestingly, a negative association was found between deadline flexibility of therapists and treatment outcome. Thus, the more lenient therapists were with homework assignment deadlines, the fewer patients improved. In a replication study of Schneider et al, the same type of therapist behaviors were found with the addition of a few more behavior categories (e.g., asking questions) (Schneider et al., 2016). However, a different distribution of the therapist behavior frequencies was found, and the outcomes were different for patients with depression than for patients with anxiety. Thus, the way online feedback is provided by therapists differs across studies, patients, interventions, and possibly also the instructions used for feedback.

1.5 Online Feedback Instructions

In addition to more general communicative behaviors of therapists, the extent to which they follow instructions for online feedback may also influence treatment effectiveness. Research on written feedback predominantly stems from the field of education. Some of the main principles can be applied to online therapeutic feedback as well. Overall, research shows that effective written feedback is timely (provided in time), selective (commenting only on 2 or 3 things that someone can change), balanced (pointing out positive aspects as well as areas in need of improvement), forward-looking (suggesting how to improve), and understandable (written in a language that someone will understand) (McKeachie and Svinicki, 2010). Instructions for training therapists in written feedback are adapted to the therapeutic process but also comparable to those used in education (e.g., beginning with a compliment, responding within 3 working days, or being careful with giving solutions). The elements in these instructions are primarily based on expert opinion rather than theory and mainly aim to motivate and support the patients, respond to the content of homework, and structure the feedback.

1.6 Study Objectives

In this study, written feedback will be studied in blended CBT, in the context of the European implementation project MasterMind (Vis et al., 2015; Mol et al., 2016; Mastermind, 2018), with a focus on therapist behaviors and on the extent to which feedback instructions are followed. In the Netherlands, iCBT for depression is slowly but increasingly adopted in routine mental health care mostly in a blended format. Blended CBT entails one integrated, standardized CBT treatment protocol that combines face-to-face sessions and digital modules to the best clinical benefit for patients and therapists (Triple-E, 2017). The evidence of blended CBT over iCBT is unfortunately still scarce. Some first studies indicate that potential benefits of blended CBT are saving therapist time without reducing therapeutic outcome, lower treatment dropout rates, more emphasis on patient self-management, more face-to-face therapy time for deepening the CBT elements, and targeting another (often more complicated) population than iCBT (Kenter et al., 2015; Schueller et al., 2017; Thase et al., 2018). The objectives of this study were to (1) identify therapist behaviors in written online communication from therapists to patients in blended CBT for adult depression in routine secondary mental health care, (2) identify the extent to which therapists adhere to feedback instructions, and (3) investigate whether therapist behaviors and therapist adherence to feedback instructions are associated with patient outcome (symptom change scores and number of completed online sessions).

2 Methods

2.1 Design

For the purpose of this observational study, the feedback messages of 45 Dutch patients that were offered blended CBT for depression by 19 therapists in routine mental health care were recruited between April 2015 and February 2017 from one outpatient clinic. This clinic was one of the participating MasterMind sites and was selected for this study because it offered a blended treatment protocol to patients within secondary health care, and the online usage information was made available for research. Patients received 219 feedback messages through a secure Web-based platform (Minddistrict, Amsterdam, The Netherlands). A mixed-method design was chosen to explore the content of the feedback messages: a directed qualitative content analysis (Hsieh and Shannon, 2005) was used to identify therapist behaviors, and a checklist for the feedback instruction adherence of therapists was developed. To explore correlations between the frequency of therapist behaviors, scores on the checklist, and patients' outcomes, an explorative quantitative approach was used.

The study was approved by a Medical Ethics Committee. They confirmed that the "Medical Research Involving Human Subjects Act" does not apply (registration number 2014.580) because the patients in this study are not required to follow certain procedures on behalf of the research (no randomization) and routine practice was followed. An internal scientific research committee approved the research proposal (CWO 2015-005).

2.2 Participants

2.2.1 Patients

Patients were recruited through their therapists. Eligible patients received study information and an information leaflet from their therapist. After approval for telephone contact with researchers for additional information, patients received an informed consent. Patients were invited for participation in MasterMind if they (1) were aged 18 years or older; (2) had a mild, moderate, or severe depression as a primary diagnosis according to the therapist; and (3) were indicated for cognitive behavioral treatment for depression following routine secondary mental health care procedures. All patients needed to explicitly consent to take part in the study. Patients were excluded from the study if they (1) did not have a valid email address and did not have a computer with Internet access and (2) did not have adequate Dutch language skills (both verbal and written).

2.2.2 Therapists

Therapists who were trained in iCBT or who were motived for iCBT were invited to participate in the MasterMind study. They were recruited through team managers and eHealth attention officers of the different therapist teams. The iCBT training consisted of a 4-hour group training, provided by the outpatient clinic. During the training, the functionalities on the online platform were shown, and therapists got the chance to practice with a fictional patient. The therapists received individual instructions, access to the blended CBT treatment protocol online, and the feedback instructions. In addition, monthly 1-hour group sessions were organized where the therapists could exchange their experiences with each other.

The feedback instructions for therapists comprised general and specific elements that go in to the structure of the messages (e.g., correct greeting, limiting to 2 subjects), readability (short sentences and paragraphs), writing style (e.g., limiting abbreviations and misspellings, use of emoticons), referring to parts of the treatment (e.g., filling in the diary, referring to the next online session), and communication skills (e.g., summarizing, not providing solutions).

2.2.3 Intervention

In the blended CBT treatment for depression of the outpatient clinic, it was agreed upon in advance that patients would receive 10 sessions online and meet with their therapist in 5 face-to-face sessions biweekly. In practice, therapists could deviate from the protocol by repeating online sessions. The online and individual face-to-face sessions were based on evidence-based treatment protocols for face-to-face CBT and are in agreement with multidisciplinary instructions for depression (Spijker et al., 2013). There were 4 core components: (1) psychoeducation, (2) cognitive restructuring, (3) behavioral activation, and (4) relapse prevention. Besides the online sessions on the treatment platform, patients were given online access to a diary and filled out questionnaires to monitor their symptoms. After each completed online session, the therapist (the same therapist as in the face-to-face sessions) wrote a feedback message to the patient. Patient and therapist could additionally communicate through a message function about practical issues (e.g., about upcoming appointments and reminders or questions about assignments).

2.3 Measures

Patient information on selected demographics (e.g., age, gender, employment status) and clinical data (e.g., use of medication) were obtained by an online self-report questionnaire at baseline. Demographic and background information (e.g., treatment and iCBT experience) of the therapists were obtained by an online self-report questionnaire at the end of the study. Usage information (e.g., number of online sessions followed and number of feedback messages) was obtained from the online platform.

Session completion was defined as the number of completed online sessions per patient. Symptom improvement was measured with the 16-item Quick Inventory of Depressive Symptomatology (QIDS) (Rush et al., 2003). The total score varies from 0 to 27, with higher scores being indicative of a higher severity of depressive symptoms. The QIDS was administered weekly on the online platform during the course of the treatment. The number of QIDS measures can vary, with up to 30 weekly measures. Of each patient, the baseline scores were included, and the last known value was used as a posttreatment score. The change score on the QIDS was calculated by subtracting the baseline measurement from the final measurement.

2.4 Coding of Therapist Behavior and Adherence

To subtract therapist behaviors from the 219 online feedback messages, a coding matrix was developed, with 9 main categories and 13 subcategories (see Appendix B1). The coding categories were based on the directed content analysis; categories from prior research (Mohr et al., 2011; Paxling et al., 2013; Holländare et al., 2016) were used to develop the initial coding scheme before analyzing the data. To score the therapists' adherence to the feedback instructions, a coding checklist was created based on the instructions that the therapists received. In total, there were 6 main categories and 19 different subcategories with a dichotomous scale (present or not present, see Appendix B2).

The coding matrix and checklist were first tested by researcher ED by coding 4 feedback messages from 2 randomly selected patients. Each of the included feedback messages was then anonymously coded and scored by researchers MM and SP. For the coding of therapist behaviors, qualitative data analysis software, ATLAS.ti 7.5.18 (ATLAS.ti Scientific Software Development GmBH, Berlin, Germany) was used.

To investigate interrater reliability, both researchers (MM and SP) coded 60 transcripts of therapeutic feedback from 10 randomly selected patients. The intraclass correlation coefficient for the therapist behaviors was .83 (95% CI 0.82-0.85) indicating good interrater reliability, based on 2-way mixed-effects agreement model (Landis and Koch, 1977; Portney and Watkins, 2000). The interrater reliability for the feedback instruction adherence categories was found to be kappa (κ) = .84 ($p < .001$, 95% CI 0.80-0.87), indicating a good agreement between the raters as well. After reaching agreement, the remaining messages (n=159) were equally divided between the 2 coders. As analysis proceeded, additional codes were developed, and the initial coding matrix and checklist were revised, discussed, and refined.

The total frequency of therapist behaviors was calculated with a query tool in ATLAS.ti. A frequency score represented the total number of times the therapist displayed a behavior in the feedback messages sent to the patient (e.g., total number of informing the patient about the assignments). To correct for the number of received feedback messages (e.g., some patients received 4 messages, and the others received 8 messages), relative frequencies were used (frequency of one category divided by the total number of frequencies of all categories per patient). The percentage of therapists' adherence to the instructions for each patient was calculated by the frequency of the adherence (e.g., the total number of times a therapist started with giving a compliment) divided by the total number of messages received by a patient.

2.5 Analyses

Statistics were conducted using IBM SPSS (SPSS Inc., Chicago IL, USA), version 22. First, descriptive statistics (means, SDs, percentages) were used to describe the patient and therapist sample, number of online sessions, and symptom improvement. Descriptive statistics were then used to examine the frequencies of therapist behaviors and percentages of therapist instruction adherence in the messages to the patients. Spearman correlation analyses, 2-sided, were conducted to assess the relationship between the therapist behaviors, feedback adherence scores and session completion, and symptom improvement. Spearman rho was used to avoid violation of assumptions of normality. Due to the small sample size, only explorative analysis, no missing values imputation techniques and no post-hoc correction for multiple testing (i.e., Bonferroni), were applied.

3 Results

3.1 Patients' and Therapists' Characteristics

A total of 45 patients (73%, 33/45 female, mean age 35.9 years) were given blended CBT in routine care by 19 therapists. Patients' characteristics can be found in Table 1. Of the 19 therapists (84%, 16/19 female), most were licensed psychologists (53%, 10/19), others were psychologists in training under supervision for health care psychologists (26%, 5/19) or mental health nurses (21%, 4/19). Moreover, 11% (2/19) of the therapists had less than 3 years of professional experience, 26% (5/19) had between 3 and 5 years of experience, 37% (7/19) had between 5 and 10 years of experience, and 21% (4/19) had more than 10 years of experience. The experience with iCBT treatments varied among the therapists: 32% (6/19) had given less than 5 iCBT treatments, 26% (5/19) had given between 5 and 10 treatments, 21% (4/19) had given between 10 and 15 treatments, and 16% (3/19) had given more than 15 treatments.

Table 1. Characteristics of patients at baseline

Patients' characteristics	n=45
Gender, female (%)	33 (73.3)
Age in years, mean (SD; range)	35.9 (12.3; 21-64)
Education level	
Secondary education level (%)	15 (36.6)
Higher education level (%)	25 (61.0)
Employment, yes (%)	21 (51.2)
Antidepressant use, yes (%)	10 (24.4)
Duration of current depression symptoms	
Duration of current depression symptoms less than 3 months (%)	8 (19.5)
Duration of current depression symptoms between 3 and 12 months (%)	22 (53.6)
Duration of current depression symptoms more than 1 year (%)	10 (24.4)

3.2 Frequencies of Therapist Behaviors and Percentages of Therapist Instruction Adherence

Appendix B1 lists the categories, definitions, and examples of the therapist behaviors. In total, 1825 therapist behaviors were coded. The most frequently used therapist behaviors were informing (27.56%, 503/1825; e.g., informing the patient about the next session or specific assignments), encouraging (23.56%, 430/1825; e.g., praising past behavior), and affirming (22.25%, 406/1825; e.g., normalizing behavior, summarizing what the patient has written or said). Making self-disclosures, confronting, and emphasizing the responsibility of the patient were never or rarely used.

An overview of the percentages of the categories, descriptions, and examples of adherence to the feedback instructions can be found in Appendix B2. The therapists adhered in most cases to correct greeting and ending of messages (95.9%, 210/219). They also scored high on writing style (93.6%, 205/219; e.g., limiting of abbreviations and misspellings) and structure (87.7%, 192/219; e.g., limiting to 2 subjects and sending the feedback within 3 working days). Therapists scored the lowest on referring (34%, 74.5/219; e.g., referring to monitoring of the symptoms or reflecting on the dairy). Within the category communication skills, therapists were very often careful with giving solutions (95.6%, 209/219) and regularly showed in their writing that they read the patients' homework (88.6%, 194/219). Formulating sentences as hypotheses is something the therapists did not often apply (10.5%, 23/219).

3.3 Session Completion and Symptom Improvement

The 45 patients completed, on average, 6.3 online sessions (Table 2). On average, patients received 4.9 feedback messages (SD 2.7; range 1-10). One feedback message contained an average of 139 words (SD 95.4; range 1-504), 14.2 words in one sentence (SD 4.0; range 1-26), 3.5 sentences in a paragraph (SD 1.9; 1-17), and 2.9 paragraphs (SD 1.6; 1-10).

From 7 patients, all QIDS data were missing because their therapists did not activate the online monitoring, leaving 38 patients for this exploration. Results on depressive symptoms showed that at baseline, the patients scored, on average, 15.8 points (SD 3.8) on the QIDS, and at post-measurement, the patients scored, on average, 11.0 points (SD 6.0), so there was an average reduction of 4.8 points (SD 6.4). Looking at symptom severity at baseline, 8% (3/38) of the patients had mild symptoms, 34% (13/38) had moderate symptoms, and 58% (22/38) had (very) severe symptoms (Table 3). At posttreatment, 21% (8/38) of the patients had no symptoms, 29% (11/38) had mild symptoms, 24% (9/38) had moderate symptoms, and 26% (10/38) had (very) severe symptoms. In total, 63% (24/38) of the patients improved on one or more categories (Table 4). Moreover, 24% (9/38) of the patients showed no change, and 13% (5/38) deteriorated in a category.

Appendix B3 contains case descriptions of 3 patients.

5

Table 2. Treatment completion and duration (n=45)

Treatment completion and duration	**n=45**
Completed online sessions, mean (SD; range)	6.3 (2.6; 2-11)
Completed face-to-face sessions, mean (SD; range)	7.1 (2.7; 2-13)
Completed face-to-face + online sessions, mean (SD; range)	13.4 (4.4; 5-23)
Treatment duration in weeks, mean (SD; range)	26.2 (11.2; 8-52)
Time period of online activity in weeks, mean (SD; range)	17.8 (10.9; 2-45)

Table 3. Severity Quick Inventory of Depressive Symptomatology scores at baseline and post-measurement

Severity QIDS	QIDS baseline n=38 (%)	QIDS post n=38 (%)
None	0 (0%)	8 (21.1%)
Mild	3 (7.9%)	11 (28.9%)
Moderate	13 (34.2%)	9 (23.7%)
Severe	20 (52.6%)	5 (13.2%)
Very severe	2 (5.3%)	5 (13.2%)

Table 4. Changes in symptom severity (n=38)

Change in depressive symptom severity	n=38 (%)
Reduction in one category	11 (28.9%)
Reduction in two categories	10 (26.3%)
Reduction in three categories	3 (7.9%)
Deterioration	5 (13.2%)
No change	9 (23.7%)

3.4 Correlations of Therapist Behaviors with Session Completion and Symptom Improvement

One correlation between therapist behaviors and session completion was found (Table 5): the therapist behavior confronting was positively correlated with online session completion (ρ = .342, p = .02). This indicates that more confrontations were related to completing more online sessions. No significant correlations were found with symptom improvement.

Table 5. Correlations of therapist behaviors with session completion and symptom improvement

Therapist behavior	**Session completion n=45**	**Change score QIDS n=38**
Emphasizing responsibility	.094	.278
Affirming	.074	.035
Clarifying the framework	.232	.069
Self-disclosure	-[a]	-
Informing	-.087	-.249
Confronting	**.342**[b]	.184
Urging	.258	.310
Encouraging	-.054	-.008
Guiding	-.055	.146
Questions	.066	.115

[a] *Indicates "not applicable"; self-disclosures did not occur*

[b] *$p < .05$, a positive correlation indicates more session completion*

5

3.5 Correlations of Therapist Instruction Adherence with Session Completion and Symptom Improvement

In Table 6, correlations of therapist instruction adherence with session completion and symptom improvement are shown. Statistically significant negative medium correlations were found between therapist instruction adherence and completed online sessions for structure ($\rho = -.340$, $p = .02$) and readability ($\rho = -.361$, $p = .02$). Meaning that the more therapists adhered to instructions containing structure (limiting to 2 subjects and sending feedback within 3 working days) and readability (short sentences and short paragraphs), the less online sessions were completed. No significant correlations were found with symptom improvement.

Table 6. Correlations of therapist instruction adherence with session completion and symptom improvement

Therapist instruction adherence	Session completion n=45	Change score QIDS n=38
Greeting/Ending	-.277	-.064
Communication skills	-.146	-.212
Structure	**-.340**[a]	-.214
Referring	.170	-.085
Readability	**-.361**[a]	-.185
Writing style	-.150	-.139

[a] *$p < .05$, negative correlations indicate less session completion*

4 Discussion

4.1 Aim of This Study

This observational study has uncovered several important factors in the content of online feedback messages in blended iCBT for depression. We further explored therapist behaviors and the extent to which therapists wrote their feedback according to their instructions. In addition, we wanted to know if therapist behaviors and adherence to the feedback instructions could be linked to patient adherence and treatment outcome. The study was carried out in a Dutch sample of participants of the MasterMind study, in routine practice, in a patient population with mild to (very) severe depressive symptoms and with a diverse group of trained and skilled therapists.

4.2 Principal Findings

Results show that therapist behaviors in relation to the online guidance are informing the patient about the functionalities on the platform, encouraging the patient by praising past behavior or inciting future behavior, and affirming by showing interest in the thoughts, emotions, and behaviors of the patient. Making self-disclosures, confronting, and emphasizing the responsibility of the patient are never or infrequently used. This is largely in line with the frequencies of the categories found by Holländare et al and may indicate that therapists use the same CBT principles in their written communication as in their face-to-face communication with the patient (Holländare et al., 2016). Previous research also found that more supportive therapists' behaviors are used frequently in iCBT and that behaviors such as confronting and self-disclosures are seldom used (Hsieh and Shannon, 2005; Paxling et al., 2013). However, in contrast to the findings of Holländare et al, we found that one and also a different therapist behavior correlated with module completion, and we also found that none of the therapist behaviors were related to symptom improvement. A possible explanation for this difference can be found in the patient group; in Holländare et al's study, patients with partially remitted depression were included within the context of a randomized controlled trial.

Although therapists applied confronting in limited cases (<1%), this was positively correlated with online session completion. In face-to-face CBT, the occurrence of confrontations has been found to be somewhat higher (6%-14.3%), but is also significantly correlated with therapy outcomes (Keijsers et al., 2000). Hill et al argued that "confrontation often interrupts the client's thinking by presenting discrepancies and another point of view [...]. Although confrontation feels negative at the time, such disruption may be a necessary foundation for change" (Hill et al., 1988).

Furthermore, therapists followed the feedback instructions that were used in this study on most of the defined elements, such as beginning with a compliment and being careful about providing solutions too soon. Different than expected, only half of the therapists formulated their sentences as hypotheses, and did so in only 10% of the feedback messages (e.g., "It sounds like you are not sure, is that correct?"). Misspellings occurred regularly: in 21.5% of the feedback messages, therapists made more than 3 spelling mistakes. One of the possible explanations for this is that the treatment platform did not contain a spelling corrector, and it may have taken therapists more time to correct their own writing. Emoticons were not used often, as only 3 therapists sometimes used an (positive) emoticon. In the "Supportive Accountability" model by Mohr and Cuijpers, it is argued that therapists may mirror the content, style, tense, and cues (e.g., emoticons) in online communication by patients to create mutual trust (Mohr et al., 2011). In this model, it is also pointed out that people pay attention to the timing and date stamps of the responses. This means responses should be timely because delays may be perceived as expressing lack of affection. In our study, the therapists sent their feedback within the limit of 3 working days in almost 80% of the cases.

Only negative associations were found with therapist instruction adherence and session completion. Providing structure and the readability was significantly negatively associated with session completion. This means that if the therapists adhered more to writing short sentences and paragraphs, and the more they limited their feedback to 2 different subjects and sent the feedback back within 3 working days, the less online sessions were completed. These findings might be explained by the adaptive, and also reactive, style of the therapists to the behavior of the patient. When patients are doing well on the online platform, they are more flexible with certain elements of the instructions. On the other hand, when patients display more difficulties or when the therapist gets the feeling that he or she is losing contact with the patient, therapists may be more inclined to adhere more to some parts of the instructions. Schneider et al also found that therapists were responsive in their online feedback and that they increased some behaviors during the course of treatment when patient depressive symptoms worsened. There are similar indications in psychotherapy, where more flexibility of therapists was found related to better treatment outcomes than therapists who were less flexible (Owen and Hilsenroth, 2014).

4.3 Strengths and Limitations

The study took place in a naturalistic setting, with routine care patients and therapists and without the restrictions of a randomized controlled trial. Patient demographic characteristics in the study sample are comparable to blended CBT research, also in routine care (Kenter et al., 2015; Thase et al., 2018). Previous studies were carried out in small samples of therapists (3-5), often trained students, who delivered treatment in a research setting (Sánchez-Ortiz et al., 2011; Paxling et al., 2013; Holländare et al., 2016; de Bruin and Meijer, 2017). With the use of a directed approach of the content analysis, the findings of the previous research were supported and extended. We found the same proportions of categories as Holländare et al with the addition of the category "asking questions" (Holländare et al., 2016). This was also found by Schneider et al when they replicated the study conducted by Paxling et al (Paxling et al., 2013; Schneider et al., 2016).

In addition, there are several limitations to this study. The generalizability of the results is limited because of the small sample size. With a greater sample size, it would have been possible to explore initial symptom severity as a predictor of the use of different therapist behaviors. The exploration of this association would be interesting for further research. Second, although this study was able to capture a group of experienced professionals, the distribution of patients over the therapist was slightly skewed. Half of the therapists treated 3 to 6 patients, and the other half treated 1 or 2 patients. Due to the small sample size, it was not possible to explore potential differences in writing style or skills between the therapists. Furthermore, in face-to-face treatment, therapist characteristics such as age, gender, and ethnicity of the therapist seemed not to be related to patient treatment outcomes (Beutler et al., 2004), but therapist facilitative interpersonal skills were found to be a successful predictor of treatment outcome (Andersson and Cuijpers, 2009). To further explore therapist online feedback, it would be interesting to look at therapeutic skills as well. In this study, there was a high variability in the number of words in the feedback messages, and it could also be interesting to further explore this. Third, within this observational study, only pre-post data and explorative analyses and, no post-hoc corrections, were used. The found correlations should be interpreted with care. Previous research showed that it is possible that therapist behaviors change over the course of treatment, with more focus on certain categories at the beginning of treatment versus the end of treatment (Holländare et al., 2016; Schneider et al., 2016). Finally, this study only focused on the online part of the blended treatment and not on the content in the face-to-face sessions.

4.4 Conclusions

In sum, this study showed that in blended CBT for depression, therapists primarily used supportive and positive communications like informing, encouraging, and affirming patient behavior. Therapists refrained from using therapeutic techniques, such as making self-disclosures, urging, and confronting. This can be explained by the way the online feedback instructions were constructed. They provided the therapists guidelines that concentrate on style and form instructions, and this is also reflected in the adherence of the therapists to most of these instructions. It can be suggested that the instructions should also focus more on “disruptive” therapeutic techniques that can foster patients to address their symptoms. The blended format can give the therapist more flexibility in writing feedback because of the combination with face-to-face contact, meaning that therapists can check the interpretation of their online feedback with the patients in the face-to-face sessions. The combination with online contact gives the therapist the possibility to incorporate elements and reflect on issues that were discussed in the face-to-face sessions. On the other hand, therapists are aware that online communications can emotionally positively and also negatively affect the patient, without them being there, and are therefore careful in their communications. The therapists may miss nonverbal cues such as facial expressions and are not able to respond immediately. Writing feedback requires the therapist to assess whether the patient can correctly understand it. The extent to which this calls for specific competencies of the “online” therapist is assumed and requires further exploration. Additional research is needed to further explore the content of online feedback. With an experimental design, more causal explanations can, e.g., be made about the amount of certain therapist behaviors, the interaction with the written content of the patients, patient expectations or the timing of feedback, and also the interaction with the contact of the face-to-face sessions. With more knowledge, instructions on feedback can be enriched, and therapists can be offered more guidance in giving feedback.

Acknowledgments

The MasterMind project was partially funded under the Information and Communications Technologies (ICT) Policy Support Programme as part of the Competitiveness and Innovation Framework Programme (CIP) by the European Community (Grant Agreement number: 621000).

CHAPTER 6

6

Does it blend? Exploring therapist fidelity in blended CBT for anxiety disorders

Geke Romijn, Simon Provoost, Neeltje Batelaan, Jeroen Koning, Anton van Balkom & Heleen Riper (submitted November 2020)

Under submission in: Internet Interventions

Abstract

Background: Blended cognitive-behavioral therapy (bCBT) combines face-to-face CBT (f2fCBT) and Internet-based CBT (iCBT) into one integrated treatment protocol, opening up new ways to deliver therapy, increase cost-effectiveness and resolve scarcity of therapist availability. When traditional therapy is transformed into a new format, there is a need to evaluate whether principles of the new protocol are consistently applied.

Methods: This study aimed to explore therapist fidelity to bCBT protocols for anxiety disorders in specialized mental healthcare and to assess whether fidelity is related to patient characteristics. Adult patients (N=44) received bCBT within a randomized controlled trial. Ratio of f2f to online sessions, therapy intensity and therapist adherence to instructions were assessed.

Results: Overall therapist fidelity with regard to ratio of blending, treatment intensity and instructions was high. Ratio of blending was significantly associated with treatment intensity (r = -.373, p = .013), meaning that treatment of higher intensity had a higher share of online sessions. Patients with more computer experience were likely to receive higher shares of online sessions (r = -.314, p = .038).

Conclusions: The blended approach was generally delivered as intended, indicating that the format is feasible in specialized mental health and supporting earlier findings on bCBT effectiveness.

1 Introduction

Cognitive-behavioral therapy (CBT) is an effective psychological treatment for anxiety disorders (Stewart and Chambless, 2009). Internet-based CBT (iCBT), where treatment is offered on an online platform, has the potential to maximize cost-effectiveness by reducing the burden of travel and reducing therapist hours. The uptake of iCBT in routine care is low, however, possibly because iCBT is not considered suitable for all patients; for example, providing web-based treatment without face-to-face contact may not be deemed acceptable for patients with severe symptoms (Gun et al., 2011). A more recent approach, blended cognitive-behavioral therapy (bCBT), combines face-to-face CBT (f2fCBT) and iCBT, partially replacing face-to-face (f2f) sessions with online sessions. Such a treatment format could address the limitations related to iCBT and may also fit better into current routine practice. In a previously conducted randomized controlled trial (RCT) (Romijn et al.) we evaluated the acceptability and effectiveness of bCBT (n=52) in comparison with f2fCBT (n=62) for anxiety disorders in specialized mental healthcare and found promising results. Patients in both groups reported high levels of treatment satisfaction, and both conditions yielded large within-group effect sizes at posttest and at one-year follow-up. A small RCT (*N*=36) of bCBT for panic disorders by another research group achieved results in line with our findings regarding acceptability and effectiveness, with medium to high effect sizes in both treatment groups and no differences in treatment satisfaction between the groups (Bruinsma et al., 2016).

Blended interventions appear increasingly popular as treatment protocols for blended therapy become more widely available (Kemmeren et al., 2016; Kleiboer et al., 2016; Kooistra et al., 2014; Nakao et al., 2018; Romijn et al., 2015). However, little is known as of yet about therapist fidelity to such protocols. Therapist fidelity is defined as the extent to which treatment is carried out as outlined in the treatment manual (Waltz et al., 1993). A lack of adequate evaluation of fidelity to blended treatment protocols can lead to incorrect conclusions about their clinical effectiveness (Prowse, 2015), because there is no way of knowing what exactly took place during the therapy. A further aim of bCBT is to improve cost-effectiveness through reduced therapist time, replacing a portion of the face-to-face sessions with online sessions. However, no clear indications for savings in terms of time or costs have been found up to now (Romijn et al.; Kenter et al., 2015; Bruinsma et al., 2016). Whether or not bCBT has been applied as designed is a key question for cost-effectiveness analysis, as it has important implications for the interpretation of cost outcomes. Suboptimal bCBT implementation may actually lead to less effective ways of treating patients. For example, in a naturalistic study by Kenter and colleagues (Kenter et al., 2015), which evaluated the use of blended treatment for anxiety and depression in routine mental health care settings, treatment time and costs increased for bCBT relative to f2fCBT, because therapists delivered the online treatment on top of the f2f sessions. The degree of treatment fidelity is important not only for understanding effectiveness and cost outcomes, but it also provides essential input for developing therapist guidelines on how to use bCBT (Bellg et al., 2004). In a recent qualitative study by Mol and colleagues (Mol et al., 2019), therapists (*N*=36) pointed to a lack of clear guidelines on incorporating online sessions as a major barrier to providing bCBT.

Protocol use in face-to-face CBT for anxiety disorders has been evaluated, and results generally indicate high fidelity to treatment protocols (Boswell et al., 2013; Zickgraf et al., 2016). In these studies, audio-taped sessions (*N*=495) (Boswell et al., 2013) or video-taped sessions (*N*=39) (Zickgraf et al., 2016) were rated in terms of therapist adherence to the CBT protocol, with mean scores emerging of 85% on a 0%-to-100% scale (SD = 10.4) and 6.18 on a 1-to-7 scale (SD = 0.51). Concerning Internet-based CBT for depression, Mol and colleagues (Mol et al., 2018) investigated adherence to feedback instructions in online sessions. They scored 219 online written feedback messages for patients with depression on a 0%-to-100% scale and concluded that therapists adhered to most of the instructions relating to issues like structure (87.7%), readability (68%), writing style (93.6%) and communication skills (69.4%). Although fidelity to f2fCBT and iCBT has thus been studied separately, no research on fidelity to bCBT has been conducted that takes both the treatment modalities of the package into account. bCBT differs from f2fCBT and iCBT in that it requires therapists not only to apply generic therapeutic skills and CBT-specific competencies in both f2f and online sessions, but also to combine the two modalities into a single integrated treatment. Insights into how therapists put this into practice are still lacking.

In the current study, we are especially interested in the ways in which therapists adhere to the blended approach – the distribution of f2f and online sessions; treatment intensity in terms of the numbers of weekly f2f or online sessions; and the instructions pertaining to the blended format, such as the explanation of the format to patients, the assisted login to the online platform during the first f2f session, and the provision of CBT-specific feedback in response to each online session. In addition, we assess the relationship between blended treatment fidelity and specific patient characteristics which therapists reportedly perceive as making patients better suited for blended therapy: younger age, employment, computer skills, higher cognitive capacities, mild-to-moderate and less complex symptoms, and preference for bCBT over f2fCBT (Mol et al., 2019).

2 Methods

2.1 Design

Data were collected within a randomized controlled trial assessing the clinical and cost-effectiveness of bCBT in comparison with f2fCBT (Romijn et al., 2015). In that trial, *N*=114 adult patients diagnosed with panic disorder, social anxiety disorder or generalized anxiety disorder were randomized to either bCBT (n=52) or f2fCBT (n=62) in one of four Dutch outpatient clinics for specialized mental health care between November 2015 and July 2017. A total of 45 patients started the bCBT treatment. Since one patient dropped out after the first session, our study of treatment fidelity analyzes data from the 44 participants who actually received bCBT. The trial was approved by the Medical Ethics Committee of the Amsterdam University Medical Centers, location VU University Medical Center (registration number 2015.073) and registered in the Netherlands Trial Register (NTR4912).

2.2 Intervention

Separate manualized bCBT protocols were developed for patients with panic disorder, social anxiety disorder and generalized anxiety disorder (Romijn et al., 2015). Their content was based on protocols for f2fCBT (Keijsers et al., 2010), which contain evidence-based elements for the treatment of anxiety disorders, such as cognitive therapy and exposure (NICE 2011; NICE 2013) (see Appendix C1). In all blended treatments, both f2f and online sessions involved therapeutic guidance by qualified psychologists. The treatment consisted of 15 weekly sessions, with 8 face-to-face sessions alternating with 7 online sessions that were followed up by scheduled online feedback from the therapist. Every course of treatment began with a f2f session. Online sessions were accessible in a secure web-based environment (Minddistrict, Amsterdam, The Netherlands). Patients and therapists accessed this platform with a personalized login. Default text templates for feedback and instructions were supplied for every online session as a therapist aid for providing feedback as intended. Therapists were free to tailor these texts to the specific needs of their clients.

The online treatment platform also offered the option of repeating an online session. Therapists could decide on that if they deemed it beneficial, for example, if the patient had not fully comprehended the content of an online session or had greatly benefited from a specific exercise in it.

Table 1 shows the protocol components for the face-to-face sessions and the online sessions. These contained instructions for the blended format, which were used to rate the extent of therapist fidelity.

Table 1. Protocol components with instructions in the blended format

Protocol component	Session	Instructions
f2f sessions		
Psychoeducation: explanation of anxiety disorder, treatment and blended approach	1	Provide explanation of treatment format: alternating f2f and online sessions Log in to online platform together with patient to provide a technical introduction
Discussing previous online session	3–15	Discuss homework and assignments from previous online session
Preparing upcoming online session	1–13	Discuss homework and content of next online session Schedule appointment for providing feedback on next online session
Online sessions		
Generic therapeutic feedback	2–14	Provide feedback containing therapist behavior that would be used in any psychotherapy intervention, such as encouraging and motivating, normalizing, empathizing, and confirming by summarizing
CBT-specific feedback	2–14	Provide CBT-driven feedback, such as helping patient identify and test automatic thoughts, helping patient identify and modify core beliefs, and helping patient plan and conduct behavioral experiments
Scheduling upcoming f2f session	2–12	Schedule an appointment for next f2f session or remind patient of the already scheduled appointment

CBT: cognitive behavioral therapy; f2f: face-to-face

2.3 Patients

Patients were invited for study participation if they (i) were aged 18 or older and (ii) met the DSM-IV criteria for panic disorder with or without agoraphobia, social anxiety disorder or generalized anxiety disorder, as diagnosed with the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I (First et al., 2002)) or the Mini-International Neuropsychiatric Interview, Plus Version (MINI-Plus (Sheehan et al., 1998; Van Vliet and De Beurs, 2007)). Exclusion criteria were (i) inadequate proficiency in Dutch, (ii) lack of e-mail address or computer with Internet access and (iii) presence of a psychotic or bipolar disorder, substance dependence or a high risk for suicide. Written informed consent was obtained from all participants before baseline assessment and randomization. A total of 44 patients received bCBT.

2.4 Therapists

Therapists were trained and experienced in delivering CBT. Prior to treating patients in the trial, all therapists received a 2-hour training course in the delivery of the blended treatment protocol, provided by researcher GR. During the training, therapists received instructions on how to apply the blended format, were shown the functionalities of the online platform, and had the chance to practice with a fictitious patient. In addition, discussion sessions were organized by the research team in which therapists could exchange their experiences with bCBT.

2.5 Measures

Information on patient characteristics – including age, gender, education level, employment status, treatment preference (bCBT vs f2fCBT), self-rated treatment motivation (0–10 scale), average weekly hours of computer use, and anxiety severity measured by the Beck Anxiety Inventory (Beck et al., 1988; Fydrich et al., 1992) – was obtained via an online self-report questionnaire at baseline. Primary diagnosis and comorbid disorders were established by a diagnostic interview (SCID-I (First et al., 2002) or MINI-Plus (Sheehan et al., 1998; Van Vliet and De Beurs, 2007)). Data on the dose and timing of f2f sessions were extracted from electronic medical records. Data on the dose and timing of online sessions were collected through the online treatment platform. f2f sessions were audio-recorded if participants consented, and these were transcribed verbatim. Feedback messages sent by therapists after online sessions were obtained from the online platform. All data were entered into Microsoft Office Excel (2016) spreadsheets.

2.5.1 Fidelity to ratio of blending

The ratio of blending for each course of treatment was calculated as the distribution of f2f and online sessions in percentages.

2.5.2 Fidelity to intensity

The bCBT protocol prescribed weekly sessions. Intensity of treatment was determined by dividing the total number of completed f2f and online sessions by the total therapy duration in weeks.

2.5.3 Fidelity to blended protocol instructions

To assess fidelity to the blended protocol in the f2f sessions and the online sessions, we developed a checklist based on the treatment protocol, containing the mandatory blended protocol components (Table 1; see Appendix C2 for the complete checklist).

Transcripts of treatment recordings and online written feedback messages were evaluated by two independent raters (GR, SP), who quantified the extent of therapist fidelity to the blended protocol instructions. One of the raters was a researcher involved in developing the treatment and conducting the trial, and one was an independent researcher not involved in the trial or in developing the treatment. Rating results were discussed until agreement was established. Interrater reliability was measured with intraclass correlation coefficients (ICC). The ICC for fidelity ratings of the f2f protocol components was .90 (p < .001, 95% CI 0.86–0.93) and for online components .90 (p < .001, 95% CI 0.88–0.91), indicating good agreement between the raters with respect to both f2f and online sessions.

2.6 Analysis

Descriptive statistics (means, standard deviations, percentages) were used to describe the sample and the observed fidelity. The relationships between fidelity scores and patient characteristics were assessed with Pearson correlations. Analyses were undertaken using SPSS, version 23 (SPSS Inc., Chicago IL, USA).

3 Results

3.1 Patient characteristics

Baseline demographics and clinical characteristics of the sample can be found in Table 2. The mean age was 36.7 years (range: 19–62 years), 53% (23/44) were female, 27% (12/44) had completed higher education and 63% (28/44) were employed.

Table 2. Patient characteristics

Patient characteristics (*N*=44)	
Age in years, mean (SD; range)	36.7 (11.0; 19–62)
Gender female, *n* (%)	23 (52)
Higher education*, *n* (%)	12 (27)
Employment, *n* (%)	28 (63.2)
Primary diagnosis, *n* (%)	
-Panic disorder	23 (52.3)
-Social anxiety disorder	11 (25.0)
-Generalized anxiety disorder	10 (22.7)
BAI score at baseline, mean (SD)	28.2 (11.6)
Comorbid disorder**, *n* (%)	25 (57)

Preference for bCBT over ftfCBT, *n* (%)	24 (55)
Treatment motivation (0–10), mean (SD)	9.3 (1.07)
Weekly hours of computer use, mean (SD)	20.5 (18.4)

BAI: Beck Anxiety Inventory; bCBT: blended cognitive-behavioral therapy; ftfCBT: face-to-face cognitive-behavioral therapy

** Bachelor's equivalent or higher*

*** Comorbid disorders: social phobia, panic disorder, agoraphobia, generalized anxiety disorder, major depressive disorder, dysthymia, posttraumatic stress disorder, obsessive-compulsive disorder, eating disorder*

3.2 Blending ratio and intensity

The mean treatment duration was 12.6 sessions (SD: 5.2), with 6.7 f2f sessions (SD: 2.6) and 6.0 online sessions (SD: 2.9) in 15.6 weeks (SD: 8.7). The mean percentage of f2f sessions in the 44 courses of blended treatment (55%) was almost equal to the prescribed 53%. Most courses of treatment (64%, n=29) contained 50% to 60% f2f sessions. In fourteen cases (32%), the share of f2f sessions was either 40% to 50% (18%, n=8) or 60% to 70% (14%, n=6); in two cases (5%) it was higher than 70% (75% and 80%).

The mean intensity of treatment was 0.89 sessions per week (SD: 0.26), slightly lower than the 1.0 sessions per week as prescribed. Five courses of treatment (11%) had a treatment intensity of exactly 1 session per week, in 13 cases (30%) treatment intensity was higher (range: 1.1-1.4) and in 26 cases (59%) treatment intensity was lower (range: 0.3-0.9).

Treatment intensity was higher (0.98 sessions/week, SD: 0.17) when the ratio of online to f2f sessions was positive than when the ratio of f2f to online sessions was positive (0.85 sessions/week, SD: 0.28). The correlation between treatment intensity and the f2f -to-online ratio was significant (r = .373, p = .013), meaning that patients who received a larger share of online sessions were likely to have higher-intensity treatment than those receiving more f2f sessions.

3.3 Fidelity to the blended protocol in face-to-face sessions

A total of 74 face-to-face sessions from 23 participants were audio-recorded. Forty-six (62%) of those sessions were conducted in full accordance with the blended instructions in the protocol, while in 28 sessions (38%) some deviations to the protocol occurred (see Table 3 for adherence to the specific protocol components).

Blended protocol instructions for the psychoeducation component, which compared to the other two components only occurred in the first session, were adhered to in 7 of 8 recorded sessions (see Appendix C3, Box 1 for an example). All therapists assisted the patient in logging into the platform during the first session to introduce the online treatment program, but in one session the therapist did not mention the alternation of f2f and online sessions in the blended treatment.

Table 3. Adherence to protocol instructions in face-to-face sessions (*N*=74 recorded sessions)

Protocol component	Protocol instruction for blended approach	Adherence	Deviations from protocol
Psychoeducation (n=8)	Explain treatment format: alternating f2f and online sessions, and login to online platform together with patient	Full adherence: 7 Partial adherence: 1 Non-adherence: 0	Therapist did not mention the alternating f2f and online sessions (1)
Discussing previous online session (n=66)	Discuss homework and assignment(s) from previous online session	Full adherence: 61 Partial adherence: 4 Non-adherence: 1	Session was mentioned, but with little or no discussion of homework and assignments (4) Previous online session was not mentioned at all (1)
Preparing upcoming online session (n=66)	Discuss homework for next online session and schedule time for feedback provision	Full adherence: 44 Partial adherence: 21 Non-adherence: 1	No time for providing feedback was scheduled (20) Homework for upcoming online session was not discussed (1) Upcoming online session was not mentioned at all (1)

f2f: face-to-face

In 5 of 66 recorded sessions (7%), therapists deviated from the protocol where the previous online session should have been discussed (Boxes 2a and 2b). In one case the online session was not mentioned at all; in other cases therapists did refer to it, but there was little or no discussion of homework and assignments.

Most deviations from the protocol occurred when the upcoming online session was to be discussed (Box 3). In 22 of 66 sessions (33%), therapists deviated from the instructions, mostly by not scheduling an appointment for providing feedback on the next online session (30%, n=20). In the other two cases, therapists did not discuss the homework for the upcoming online session.

3.4 Fidelity to the blended protocol in online sessions

Therapists provided a total of 257 feedback messages on 257 online sessions (see Table 4 for adherence to specific protocol components). In 167 messages (65%), blended protocol instructions were fully adhered to, meaning that the therapist had provided both generic therapeutic and CBT-specific feedback, and had scheduled the appointment for the next f2f session (see Appendix C3, Boxes 4a and 4b for examples). Generic therapeutic feedback was provided in 232 messages (90%), CBT-specific feedback in 184 messages (72%) and an appointment for the upcoming f2f session was scheduled in 208 messages (81%).

Thirty-four (13%) of the 257 online sessions were repeated. Feedback messages on repeated sessions were usually short and practical in nature (Boxes 5a to 5c) and usually did not contain CBT-specific feedback (31 of 34 messages).

Table 4. Adherence to protocol instructions in online sessions (*N*=257 sessions)

Protocol component	Online sessions with full adherence to instructions
Generic therapeutic feedback	n=232 (90%)
CBT-specific feedback	n=184 (72%)
Scheduling upcoming f2f session	n=208 (81%)

CBT: cognitive behavioral therapy; f2f: face-to-face

3.5 Correlations of patient characteristics with treatment fidelity outcomes

6

Table 5 shows correlation between patient characteristics and fidelity outcomes. There was a significant association between the ratio of f2f to online sessions and a patient's experience with the use of computers (r = −.314, p = .038), meaning that the treatment of patients more experienced with computers was likely to contain a larger percentage of online sessions. No significant associations between other patient characteristics and fidelity outcomes were found.

Table 5. Correlations of patient characteristics with treatment fidelity outcomes (N=44, Pearson's r)

Patient characteristics	Blending ratio	Intensity	f2f fidelity score (n=23)	Online fidelity score
Age	.213	.163	–.083	.244
Higher education	–.104	.159	–.340	–.111
Employed	.016	.130	–.110	–.157
Baseline BAI score	.238	–.290	–.024	.114
Comorbid disorder	–.227	–.023	.080	–.275
Preference for bCBT	.161	.092	.228	–.132
Weekly computer hours	–.314*	.140	.024	–.159

f2f: face-to-face; BAI: Beck Anxiety Inventory; bCBT: blended cognitive-behavioral therapy

* $p < .05$

4 Discussion

4.1 Principal findings

The aim of this paper was to explore therapist fidelity to blended cognitive-behavioral therapy (bCBT) protocols for anxiety disorders in specialized mental health care. Additionally, we wanted to gauge the influence of patient characteristics on bCBT fidelity, since therapists believe some patients are better suited for bCBT than others (Mol et al., 2019).

Overall, therapist adherence to the instructions in the blended treatment protocol was high. The mean intensity of treatment was 0.89 sessions per week (SD: 0.26), slightly lower than the 1.0 sessions per week as prescribed. The ratio of f2f to online sessions was associated with treatment intensity – a possible indication that a larger share of online sessions enables a higher intensity of treatment. That seems relevant in the light of meta-analytic findings by Cuijpers and colleagues (Cuijpers et al., 2013b) showing the importance of treatment intensity: they found that an increase from one to two sessions per week in psychotherapy for depression boosted the effect size g by 0.45, with the total number of sessions held constant.

Our inspection of patient characteristics showed a significant association between the ratio of blending and patients' experience with the use of computers, indicating that those with more computer experience were more likely to receive a higher share of online sessions. Other patient characteristics, such as pre-treatment anxiety severity or comorbidity, were not associated with fidelity outcomes. This finding is in line with previous findings for f2f therapy (Boswell et al., 2013; Zickgraf et al., 2016) and refutes therapists' believe that patients with mild-to-moderate and less complex symptoms are better suited for bCBT (Mol et al., 2019).

4.2 Strengths and limitations

To the best of our knowledge, this study was the first investigation of therapist fidelity to a blended treatment protocol that assessed both the face-to-face and the online elements of the treatment. This enabled us to examine what actually happened during the blended treatment: did it blend? In other words: did therapists adhere to the blended approach in terms of the distribution of f2f and online sessions; treatment intensity and the instructions pertaining to the blended format.

Evaluating fidelity is a time-consuming process, and it becomes even more complex when two treatment modalities are integrated into one treatment protocol. For this reason, treatment fidelity is often not examined in intervention studies, and had not yet been evaluated for blended treatment at all, even though it is essential to the interpretation of treatment outcomes and to successful implementation of a blended format.

A limitation of the present study is that recordings of f2f sessions were not available for all participants and all sessions. Only 23 of the 44 participants consented to audio recording, and in other cases recording errors occurred or therapists forgot to record sessions. Possibly, patients and therapists with more positive beliefs about the blended treatment format were more motivated to help collect audio recordings. That may have biased our findings.

Comparing our results to other studies might be complicated by the difference in use of definitions. We operationalized treatment fidelity as adherence by therapists to the specific blended approach as prescribed in the treatment protocol of the current trial. Other definitions of blended treatment (Wentzel et al., 2016) or of therapist fidelity (Carroll et al., 2007; Gearing et al., 2011) may be used in other studies.

4.3 Clinical and research implications

Opportunities to improve therapist fidelity to the blended treatment format appear to lie in enhancing therapist recognition of f2f and online sessions as equally important elements of treatment. If that is not acknowledged, online sessions cannot adequately replace f2f sessions. In the current study, therapists often failed to schedule appointments for providing feedback on online sessions, which could be an indication that a therapist sees those sessions as merely supportive to the f2f sessions and this idea can unconsciously be transferred to the patient. This requires attention in the training of therapists. Furthermore, if an appointment calendar function were added to the online platform, that might improve fidelity to this protocol component and heighten therapists' awareness of the importance of online elements in blended treatment.

Previously, a lack of clear guidelines has been identified as a barrier to the use of bCBT (Titzler et al., 2018; Mol et al., 2019). In some cases in the current study, treatment intensity was higher than the intended one session per week, and that higher intensity was sometimes caused by a lack of clarity about how to integrate the online element into the treatment. This points to the need for clear instructions about online communication (such as how to deal with flexible, on-demand online contact opportunities and how much therapist time is available for online activities) and to the necessity of more intensive therapist training, which can prevent bCBT from becoming too demanding for therapists or too costly.

One benefit of a blended format, as found in earlier studies, is that it can enhance therapists' adherence to the treatment protocol (Titzler et al., 2018; Mol et al., 2019). In the current study, we indeed found high therapist fidelity in most courses of treatment. In the cases with lower fidelity, however, our fidelity results showed quite some variability in ratios of blending and in treatment intensity, indicating that some therapists may feel the need for a less fixed protocol. Dosing flexibility could, for example, be accomplished by including a number of mandatory sessions and a number of complementary sessions. The flexible character of online sessions facilitates the option of shortening or expanding treatment. Offering a customizable blended protocol based on therapist and patient preferences has been suggested before (van der Vaart et al., 2014; Wentzel et al., 2016; Titzler et al., 2018; Kemmeren et al., 2019), and future research should further explore this option and investigate what degree of flexibility might be feasible.

Finally, an interesting topic for subsequent research would be whether therapist variables are associated with fidelity outcomes. Identifying therapist characteristics that predict fidelity to bCBT could assist mental health care services in the selection and training of professionals.

4.4 Conclusions

Our findings showed that the blended treatment was generally conducted as intended, indicating that implementation of bCBT in routine care settings is feasible. These results also confirm the internal validity and enhance the external validity of the randomized controlled trial from which the data derive (Romijn et al., 2015).

The current study was conducted prior to the coronavirus crisis. The outbreak of a pandemic disease highlights the relevance of online treatment as an important element of routine care practice (Wind et al., 2020). Blended interventions are likely to be of critical importance in post-corona mental health care.

CHAPTER 7

7

Improving adherence to an online intervention for low mood with a virtual coach: Study protocol of a pilot randomized controlled trial

Simon Provoost, Annet Kleiboer, José Ornelas, Tibor Bosse, Jeroen Ruwaard, Artur Rocha, Pim Cuijpers & Heleen Riper (2020)

BMC Trials, 21, 860
doi: 10.1186/s13063-020-04777-2

Abstract

Background: Internet-based cognitive-behavioral therapy (iCBT) is more effective when it is guided by human support than when it is unguided. This may be attributable to higher adherence rates that result from a positive effect of the accompanying support on motivation and on engagement with the intervention. This protocol presents the design of a pilot randomized controlled trial that aims to start bridging the gap between guided and unguided interventions. It will test an intervention that includes automated support delivered by an embodied conversational agent (ECA) in the form of a virtual coach.

Methods/design: The study will employ a pilot two-armed randomized controlled trial design. The primary outcomes of the trial will be (1) the effectiveness of iCBT, as supported by a virtual coach, in terms of improved intervention adherence in comparison with unguided iCBT; and (2) the feasibility of a future, larger-scale trial in terms of recruitment, acceptability, and sample size calculation. Secondary aims will be to assess the virtual coach's effect on motivation, users' perceptions of the virtual coach, and the general feasibility of the intervention as supported by a virtual coach. We will recruit N=70 participants from the general population who wish to learn how they can improve their mood by using Moodbuster Lite, a 4-week cognitive-behavioral therapy course. Candidates with symptoms of moderate to severe depression will be excluded from study participation. Included participants will be randomized in a 1:1 ratio to either (1) Moodbuster Lite with automated support delivered by a virtual coach, or (2) Moodbuster Lite without automated support. Assessments will be taken at baseline and post-study four weeks later.

Discussion: The study will assess the preliminary effectiveness of a virtual coach in improving adherence and will determine the feasibility of a larger-scale RCT. It could represent a significant step in bridging the gap between guided and unguided iCBT interventions.

1 Background

The most widely studied online interventions for depression are those based on cognitive-behavioral therapy (CBT) (Cuijpers et al., 2013a). Such interventions may be guided or unguided. Guided interventions typically include regular feedback and support by professional health care workers, licensed therapists or trained volunteers, either via secured email exchange or via messaging systems within the intervention platforms. In shorter interventions, mostly up to eight sessions, support often takes the form of coaching, but in more intensive types of treatment it may be more therapeutic in nature. Guided interventions have been found more effective in terms of symptom improvement (Spek et al., 2007; Johansson and Andersson, 2012; Richards and Richardson, 2012; Karyotaki et al., 2019). That may be explained by a more positive effect of the guidance on motivation and engagement, and hence on adherence rates (Mohr et al., 2011; Kelders, 2015). However, as guided interventions require the involvement of supportive humans, unguided interventions are potentially more scalable, more accessible, and less expensive (Riper et al., 2010). This study is part of a project to bridge the gap between guided and unguided self-help internet-based CBT (iCBT) interventions for depression, using embodied conversational agents (ECAs) to automate coaching support. ECAs are more or less autonomous and intelligent software entities with a graphical embodiment. They are used to communicate with the user (Ruttkay and Pelachaud, 2004).

The idea of using ECAs in psychological treatment procedures goes back roughly a decade (Bickmore and Gruber, 2010), and a recent scoping review has shown that many different such applications have since been developed for a variety of common mental health disorders (Provoost et al., 2017). In the context of depression, ECAs have been proposed for a broad range of applications. For example, ECAs have taken on the role of an interviewer that engages in face-to-face interaction with users to make them feel more comfortable in talking about and sharing distressing information (Devault et al., 2014), or the role of a virtual nurse who guides hospital patients with depression through their discharge procedure (Bickmore et al., 2010b), or that of an empathic therapist who helps people navigate the Beck Depression Inventory questionnaire (Pontier and Siddiqui, 2008). A number of studies have applied ECAs in the context of an iCBT for depression. Study designs varied widely. Martínez-Miranda and colleagues conducted a pilot study in which an ECA supported users throughout a CBT intervention (Martínez-Miranda et al., 2014). Their evaluation, involving N=8 adult participants with mild to moderate depression, focused primarily on the feasibility of the cognitive change model employed by the ECA to regulate its own emotional responses, for example, by providing more empathic feedback or facial expressions. In a randomized controlled study by Kelders of an online acceptance and commitment therapy involving N=134 adults with mild to moderate depression, half of the participants received automated feedback accompanied by a picture of a clinician and the other half received human support (Kelders, 2015). The study concluded that, although participants receiving human support were more involved in the intervention than those receiving automated feedback (as scored on the Personal Involvement Inventory (Zaichkowsky, 1994)), they were not significantly more adherent in terms of intervention completion. A pilot study by Ring and colleagues aimed to create a one-on-one therapeutic conversation with a virtual counselor (Ring et al., 2016). In a pre–post-test study design including N=10 participants with mild to moderate depression, most users reported that the agent understood their emotions, but no significant improvements in depressive symptoms were found. Another pre–post-test pilot study investigated the acceptability and usability of a user-adapted, ECA-supported interactive platform addressing depression and suicide symptoms in a convenience sample of N=60 participants (Bresó et al., 2016). It concluded that system usability and the acceptability of the agent's emotional responses were sufficient for the researchers to continue preparing the system for an initial clinical trial. A study by Fitzpatrick and colleagues looked at the feasibility, acceptability, and preliminary effectiveness of a conversational agent called Woebot, which delivered CBT-based self-help content in a text-based conversational format (Fitzpatrick et al., 2017); N=70 university students who self-identified as experiencing depression and anxiety symptoms were randomized to using Woebot or to reading a book on depression. The intervention group reported significant reductions in depressive symptoms compared with the control group (d = 0.44). In another study, Suganuma and colleagues investigated the feasibility and acceptability of an ECA-delivered CBT-based intervention that aimed to determine users' mental and physical status in order to make appropriate behavioral suggestions. A non-clinical intervention group of n=191 users was compared with n=263 study participants who did not use the intervention. The intervention showed some initial effectiveness in terms of mental health improvement (Suganuma et al., 2018). Many of the applications described in this paragraph were judged acceptable and feasible, and some of the studies even showed that positive treatment effects can be accomplished using ECA-based interventions (e.g., (Fitzpatrick et al., 2017; Suganuma et al., 2018)).

Although the studies just reviewed have shown promising results, most did not focus on ECAs in a supportive role as an adjunct to an iCBT intervention (intervention + ECA), but rather on the ECA as a medium through which iCBT could be delivered (intervention = ECA). In order to strengthen the evidence for the use of ECAs as an adjunct to improve iCBT interventions, a study would need to compare an ECA-supported intervention with the same intervention with either human support or no support. Of the studies cited above, only the one by Kelders (Kelders, 2015) used such a design. That study, however, focused primarily on automated support through text messages, with the support embodied with a picture of a clinician. Though this does satisfy our criteria for what an ECA is, we might question how well the results generalize to interventions utilizing more sophisticated ECA technology. We aim to address this gap in the literature by comparing outcomes of participants in an existing intervention with added ECA support (our intervention group) with the outcomes of participants in the same intervention without ECA support (our control group). Our general hypothesis is that by simulating a number of human support factors—specific factors such as motivational interviewing techniques and feedback to CBT exercises and common factors such as empathic communication (Wampold, 2015)—an ECA can positively affect motivation and engagement, and thereby adherence rates. This, in turn, may increase the clinical effectiveness of iCBT interventions in which traditional human support is unavailable (Donkin et al., 2011). Given the novelty of our approach, which combines an existing iCBT intervention with ECA support, we have opted for a pilot randomized controlled trial, whose primary aims will be to compare adherence rates between the two study groups and to assess the feasibility of a future larger-scale trial. Secondary aims include assessing within and between-group participant motivation for performing and continuing the intervention, gauging users' acceptance of and perceived relationship with the supportive ECA, and estimating the feasibility of the entire system in terms of user satisfaction, usability, and preliminary effectiveness.

2 Methods/design

2.1 Study design

The study is designed as a pilot non-blinded two-armed randomized controlled trial (*N*=70) in which people with low mood from the general population will be randomly allocated either to an intervention for improving mood with automated support delivered by a virtual coach (n=35) or to the same intervention without the automated support (n=35). The study protocol has been approved by the Medical Ethics Committee of the VU University Medical Centre, Amsterdam (registration number 2019.388). Written informed consent will be obtained from all participants. Figure 1 displays the flowchart of the study design in accordance with the SPIRIT guidelines (Moher et al., 2010; Schulz et al., 2010).

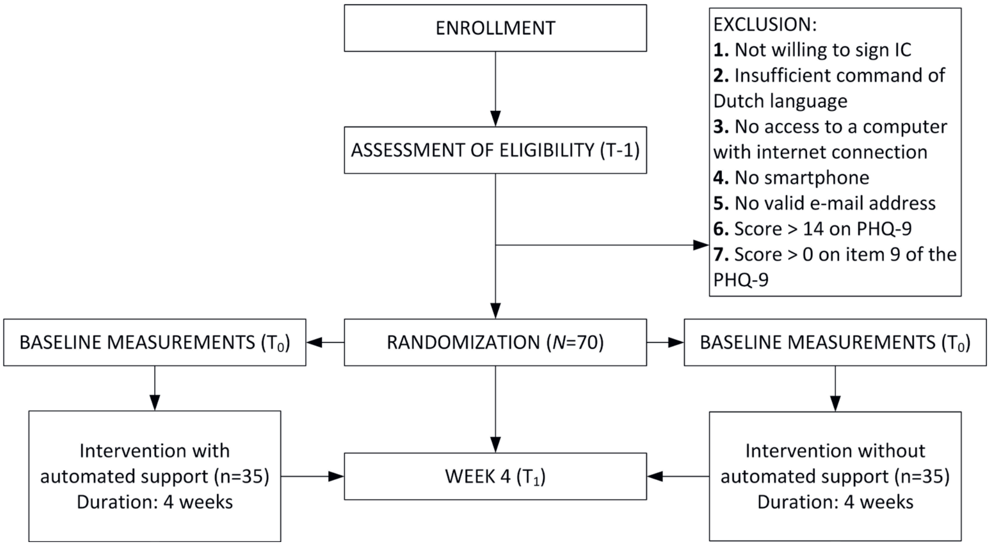

Figure 1. Flowchart of the study design

2.2 Assessments

Assessments will be taken after enrollment (T–1), at baseline (T0), and at the end of study participation four weeks after baseline (T1). Questionnaires will be self-administered and completed online. Table 1 provides an overview of the measures employed at specific time points.

Table 1. Measures administered at each assessment interval

Questionnaire	Aim	Enrollment (T–1)	Allocation (T0)	Post-study (T1)
Mental health				
PHQ-9	Screener	X		
HADS-D	Mental health		X	X
Feasibility				
SUS	System usability			X
CSQ-I	User satisfaction			X
Study completion	Reasons for non-adherence			X
Motivation				
SMFL	Current use		X	X

Continued use	Continued use	X
Coach acceptance		
WAI-SR*	Relationship	X
Acceptance*	Acceptance	X

** Applies to intervention group only.*

2.3 Participants

2.3.1 Inclusion criteria

People from the general population in the Netherlands, aged 18 years or older, will be eligible for recruitment if they express a desire to learn how to improve their mood.

2.3.2 Exclusion criteria

Candidates will be excluded from the study if they (i) are not willing to sign the informed consent form, (ii) do not have adequate proficiency in the Dutch language, (iii) do not have a computer with internet access, (iv) do not have a smartphone, (v) do not have a valid e-mail address, (vi) have moderate to severe depression, or (vii) are identified as at risk for suicide. The Patient Health Questionnaire 9 (PHQ-9) will be used to assess whether exclusion criteria vi and vii apply. Excluded candidates will receive an email detailing the reason for their exclusion. If exclusion criterion vi applies (a score of 15 or higher on the PHQ-9), they will be advised to contact their general practitioner; and if vii applies (a score of 1 or higher on PHQ-9 item 9), they will also be referred to a national help and crisis line for people at risk of suicide (https://www.113.nl).

2.4 Recruitment

Participants will be recruited in an open recruitment strategy via advertisements in digital media (Facebook, Google Ads) and http://www.link2trials.com. Interested persons will be invited to express their interest in participation by filling out a web form, after which they will receive an information brochure and an informed consent form. People who sign the consent form will receive a link to the online screening questionnaire and, once found eligible for participation, will be sent final instructions and login credentials for taking part in the study. Participants will receive 30 euro if they complete the T1 assessments, irrespective of how much time they have committed to the course. They will be free to discontinue study participation at any time, and participation places no restrictions on their use of alternative sources of help.

2.5 Randomization and blinding

Participants will be randomly assigned by an independent researcher to either Moodbuster Lite with automated support (intervention group) or Moodbuster Lite without automated support (control group). That will take place in a 1:1 ratio and on the basis of a computer-generated block randomization table with random block sizes (Efird, 2010). Group allocation cannot be blinded to participants, because a description of the study's research aim—improving intervention adherence with automated support by a virtual coach—must be provided in the information letter; whether or not automated feedback is provided will hence be obvious to participants. The principal investigator, who coordinates the study and conducts the data analysis, will not be blinded to the participants' group allocation.

2.6 Interventions

2.6.1 Moodbuster Lite

Moodbuster Lite is a four-week therapeutic course aimed at improving mood. It is a light-weight version of the Moodbuster for Depression intervention (Warmerdam et al., 2012; Kleiboer et al., 2016) and consists of a web-based and a mobile component. Compared to Moodbuster for depression, which also contains a number of cognitive therapy-based modules, the focus of Moodbuster Lite is on behavioral activation (Lewinsohn et al., 1976). Through activity scheduling, participants learn to turn a 'negative spiral', with few pleasant activities leading to few positive stimuli, a low mood, and little incentive to perform more activities, into a 'positive spiral', with more pleasant activities leading to more positive stimuli, a better mood, and incentive to remain active. A secure web-based platform provides access to online lessons, homework exercises, a mood graph, and a calendar. A smartphone application, designed for both Android and iOS, prompts participants three times a day with a request to rate their current mood, and an overview of the participant's responses is shown in both the app and the web platform's mood graph. The course consists of three lessons that were adapted from the Moodbuster for Depression intervention to fit the low-mood context of this study: (1) Introduction, (2) Psychoeducation, and (3) Pleasant Activities. The first lesson has also been extended with some exercises based on motivational interviewing (Rollnick et al., 2008) to increase participants' motivation for completing the course. For the purpose of the current study, an optional virtual coach has been embedded into the platform to provide automated support at the beginning and the end of every lesson and halfway through Lesson 3, the longest lesson. For this study, participants are advised, but not obliged, to complete the intervention in a time span of four weeks. On completion, participants retain their access to the platform for about another five months. An overview of the intervention is shown in Table 2.

Table 2. Overview of the Moodbuster Lite course as used in this pilot RCT

	Lesson 1	Lesson 2	Lesson 3
Topic	Introduction	Psychoeducation	Pleasant Activities
Time frame	Week 1	Week 1	Weeks 2–4
Length	22 pages	11 pages	17 pages
Conversations with virtual coach	2	1–2	2–3

2.6.2 Automated support

2.6.2.1 Technical implementation

Automated support is delivered by a virtual coach in the form of an ECA. The ECA has been implemented in TyranoBuilder (TyranoBuilder, 2018), a JavaScript-based software package for the development of visual novels that can be used to implement text-based dialogues with a virtual character. Our choice for TyranoBuilder was strongly motivated by the fact that applications can be exported in a browser-format that allows them to be embedded in web pages (Figure 2).

Figure 2. The virtual coach embedded in the Moodbuster Lite platform

2.6.2.2 Embodiment

We have embodied the ECA using a single two-dimensional static cartoon-like character, taking into account the following recommendations from the literature on ECAs for motivational and coaching purposes. We have opted for a cartoon-like embodiment, as increased realism is not that important for involvement, distance, and use intentions, and may even set high expectations that the ECA cannot meet (van Vugt et al., 2009). With regard to gender, we have chosen a female embodiment, as that is what people on average prefer (Canidate and Hart, 2017). The ECA is endowed with a number of facial expressions (friendly, smiling, compassionate, questioning; see Figure 3), such that it can convey a sense of empathy (Baylor and Kim, 2009); we have not given the ECA negative facial expressions (Pagliari et al., 2012). Finally, the ECA is designed to look as if it could be part of a therapy team, increasing its credibility by giving it a semi-formal friendly appearance and placing it before a background reminiscent of a therapy office (Baylor, 2011).

Figure 3. The four different expressions of the virtual coach: friendly, smiling, compassionate, questioning (left to right)

2.6.2.3 Conversations

The conversations have been designed in collaboration with a licensed therapist, and are based on guidelines for e-coaching (Mol et al., 2018) and principles of motivational interviewing (Rollnick et al., 2008). Some examples of guidelines for providing feedback we have applied are to (1) use correct greetings and closings; (2) use communication skills such as beginning a message with a compliment; (3) structure feedback, for example by not giving feedback on more than two subjects; (4) refer to things the participants have done, such as completing exercises or recording their moods; and (5) keep text readable by using short, clear sentences. With regard to the motivational interviewing, we have focused on increasing an individual's willingness to change behavior, as well as on their confidence in their ability to do so, both of which are important for being "ready" to change. Baseline values of a participant's willingness to change and confidence in their ability to do so are established using the importance and confidence ruler exercises in Lesson 1 of the intervention. If importance or confidence is low, the virtual coach presents specific exercises aimed either at increasing the discrepancy between a participant's goals and their current behavior and emphasizing the importance of change, or at enhancing a participant's self-efficacy and emphasizing confidence in their ability to change. These elements have been incorporated into all the conversations except the introductory and final ones, thus providing us with the general conversation structure shown in Table 3. Conversations after each lesson always take place, focused on providing feedback, while conversations before a lesson take place only if motivation is considered low or if the previous lesson received a negative evaluation. Such evaluations can be given by free text input at the end of each lesson, and a sentiment analysis algorithm (Provoost et al., 2019) is used to determine its valence (negative or positive).

Table 3. The differential stages in the conversations

Stage	Before the lesson	After the lesson
1	Greeting	Greeting
2	Compliment or positive note	Compliment or positive note
3	Reflection on evaluation of previous lesson	Reflection on current lesson
4	Re-evaluation of confidence or willingness	Confidence or importance work
5	Reference to current lesson	Reference to next lesson
6	Goodbye	Goodbye

2.6.2.4 Conversation trees

The conversations take place through text-based messages appearing beneath the virtual coach (see Figure 2), and the user can proceed through the conversations by clicking the mouse button, or now and then by selecting or providing an answer when asked a question. Although much progress is currently being made in speech and natural language processing, we decided to represent our dialogues in textual conversation trees for several reasons: (1) speech and natural language processing are still far from flawless; (2) automatic interpretation and accurate response to semantic content is difficult; (3) conversation trees can be more easily interpreted by domain experts such as clinical psychologists; (4) conversation trees are deterministic, meaning that there is an exhaustive set of possible conversations that can be checked for inconsistencies; and (5) certain paths through the tree can be made conditional, for example based on an answer to an earlier question in the lesson or conversation, thus enabling conversations to be personalized. For illustrative purposes, Figure 4 shows an excerpt from one of the conversation trees. The diamond represents a decision point in the conversation tree, rectangles represent utterances by the virtual coach, and circles indicate the corresponding facial expressions.

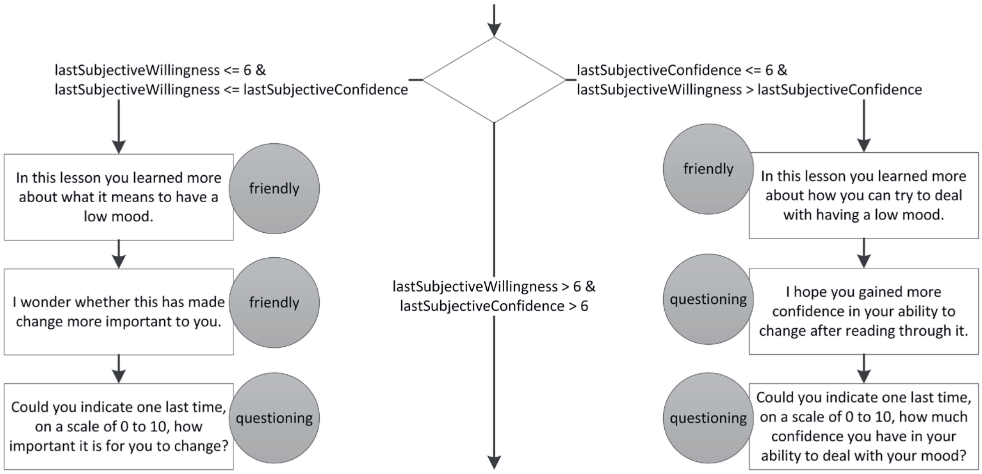

Figure 4. A conversation tree snippet from the dialogue that takes place after the second lesson

The excerpt compares the latest confidence and willingness ratings provided by the user. If both values are higher than 6, the confidence and importance work is skipped. If one value is 6 or lower, the user is asked to re-evaluate the lower rating, prioritizing willingness over confidence, after which the tree continues with a suitable exercise. Table 4 provides additional information about the variables used in this excerpt.

Table 4. Additional information about the variables used in the conversation tree excerpt depicted in Figure 4

Variable	Origin	Use
name: lastSubjectiveWillingness value range: 0–10	The latest "willingness to change" rating provided by the user, during either the first lesson or a subsequent conversation	Determines whether importance work is needed, and is compared with the confidence rating to determine which of the two to prioritize
name: lastSubjectiveConfidence value range: 0–10	The latest "confidence in ability to change" rating provided by the user, during either the first lesson or a subsequent conversation	Determines whether confidence work is needed, and is compared with the willingness rating to determine which of the two to prioritize

7

2.7 Trial Organization

The study is run from VU University Amsterdam, with no other study centers participating. The principal investigator is responsible for coordinating the study, which includes the recruitment of participants and the informed consent procedure, responding to questions and requests from (potential) participants, providing participants with access to the study materials, monitoring participants throughout the study, handling participant reimbursements, data collection, and reporting on the progress of the study to the steering committee members and medical ethical committee. The steering committee (see title page for members) agreed on the final version of this protocol, and is responsible for reviewing the progress of the study, and for agreeing on changes to the protocol or study materials, if necessary, to keep the study running properly. Meetings of the steering committee are scheduled when necessary. The trial management committee is composed of the principal investigator, and project leader. It is responsible for the study planning, organization of steering committee meetings, reporting to the medical ethical committee of study progress, maintenance of the trial master file, budget administration, and data verification. The trial management committee meets on a monthly basis. An IT team is responsible for the maintenance of the intervention platform and data collection from the platform. As this is a relatively small pilot study there is no Stakeholder and Public Involvement Group.

Earlier large-scale research using the Moodbuster platform did not result in any known Serious Adverse Events (SEAs) or Serious Adverse Device Events (SEDAs). If SEAs or SEDAs do occur, they will be discussed in the research team and reported to the Dutch Health and Youth Care Inspectorate. Any other adverse events reported spontaneously by the participants or observed by the investigators will be recorded. Due to the low-risk nature of the study there is no anticipated harm and compensation for trial participation. Participants can contact an independent researcher if they run into issues during the study, and a licensed psychiatrist can be consulted in case issues of a medical or mental health related nature arise.

Significant amendments to the study protocol will be communicated to the medical ethical committee that approved the study, and an update will be made to the study information in the Dutch Trial Registry. Results will be published in a peer-reviewed journal, and reported to the medical ethical committee that approved the study.

2.8 Primary outcome measures

2.8.1 Adherence

The primary outcome measure will be intervention adherence. According to the definition we have adopted, “adherence” describes the extent to which individuals are exposed to the content of the intervention (Christensen et al., 2009). Previously this has been operationalized by dividing the number of completed sessions or modules by the maximum number (van Ballegooijen et al., 2014), but because our 3-lesson course is relatively short, we will use the completed and maximum numbers of pages that make up the lessons. Including conversations with the coach, Lesson 1 has 22 pages (20 in the control condition), Lesson 2 has 13 (11 in the control condition), and Lesson 3 has 20 pages (17 in the control condition). As a secondary way of measuring adherence, we will look at the ecological momentary assessment of mood via the smartphone application, whereby (similarly to adherence to the intervention content) we will operationalize adherence as the number of mood assessments made divided by the maximum possible number. There will be three mood assessments every day, meaning that participants can answer a maximum of 84 mood rating requests during the four weeks of the study.

2.9 Secondary outcome measures

2.9.1 Motivation

Motivation for taking part in the intervention will be assessed in both groups by the Short Motivation Feedback List (SMFL) (Jochems, 2016). It consists of eight 10-point Likert-scale items ranging from "completely disagree" to "completely agree", designed to capture the level and type (external, introjected, or identified) of a patient's treatment motivation. The SMFL is based on self-determination theory (Ryan and Deci, 2008) and has been found to have a congeneric reliability ranging from 0.81 to 0.93 (Jochems, 2016). There are two different versions. The pre-intervention version will be assessed at baseline (T0) and the post-intervention version after four weeks (T1). Motivation to continue using the intervention will be assessed by a single statement, "I intend to continue using the platform to schedule and perform activities," assessed on a 5-point Likert scale ranging from "completely disagree" to "completely agree."

2.9.2 Relationship with the coach

After study completion (T1), participants in the intervention group will assess their relationship with the virtual coach on the Bond scale of the Revised Short Version of the Working Alliance Inventory (WAI-SR) (Hatcher and Gillaspy, 2006; Stinckens et al., 2009). The WAI-SR rates the quality of the therapeutic relationship with the virtual coach, and it has been adjusted to our context by replacing the name of the therapist with the word "coach." The Bond scale consists of four 5-point Likert-scale items ranging from 1 (seldom) to 5 (always). The final raw score may range from 4 to 20, with higher scores indicating a better bond between participant and coach. The psychometric properties of the questionnaire are satisfactory (Hatcher and Gillaspy, 2006).

2.9.3 Acceptance of the coach

Acceptance of the virtual coach will be assessed in the intervention group after four weeks (T1) using a set of six 7-point Likert-scale items. This scale has been previously used to measure attitudes toward a virtual discharge nurse (Bickmore et al., 2010b) and has been adjusted to our context of iCBT. An overview of the items is provided in Table 5. Participants are asked to elaborate on their answers to each of these questions in an open text format.

Table 5. Self-report measures of attitudes toward the virtual coach

Measure	Question	Likert-scale extremes
Satisfaction	How satisfied were you with the virtual coach?	Not at all – Very satisfied
Usability	How easy was it talking to the virtual coach?	Easy – Difficult
Continue	How much would you like to continue working with the virtual coach if the course continued?	Not at all – Very much
Relationship	How would you characterize your relationship with the virtual coach?	Complete stranger – Close friend
Preference	Would you rather have followed the course with or without the virtual coach?	Definitely prefer no coach – Definitely prefer virtual coach
Adherence	How likely is it that you will follow the virtual coach's advice?	Not at all likely – Very likely

2.9.4 System usability

Usability of the platform will be assessed after week 4 (T1) by the System Usability Scale (SUS) (Brooke, 1996). The SUS is composed of ten 5-point Likert-scale items with response options ranging from 0 (strongly disagree) to 4 (strongly agree). Total scores are converted to a scale ranging from 0 to 100, where higher scores are indicative of higher platform usability. The SUS is considered a reliable instrument, and scores higher than 68 indicate "good" usability (Bangor et al., 2008).

2.9.5 User satisfaction

User satisfaction with the web-based intervention will be assessed by the Client Satisfaction Questionnaire for internet-based interventions (CSQ-I) (Boß et al., 2016), an adaptation of the original CSQ (Larsen et al., 1979). The CSQ-I is composed of eight 4-point Likert-scale items with response options ranging from "does not apply to me" to "applies to me." Total scores range from 8 to 32, with higher scores indicating greater client satisfaction. The CSQ-I has been found to be a reliable instrument (Boß et al., 2016).

2.9.6 Mental health status

Mental health status will be assessed using the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) (Snaith and Zigmond, 1986), consisting of seven items, each assessed on a 3-point scale. Total scores range from 0 to 21, and higher scores indicate more severe depression symptoms. An often-used cut-off score for the HADS-D is 8 or higher, standing for "relevant symptoms of depression." The HADS has been shown to be a reliable and valid instrument in various populations (Spinhoven et al., 1997).

2.9.7 Mood

Participants' mood will be assessed through ecological momentary assessments on a smartphone application that works on both Android and iOS systems. The application prompts participants three times a day to rate their mood on a scale of 1 to 7 (see Figure 5).

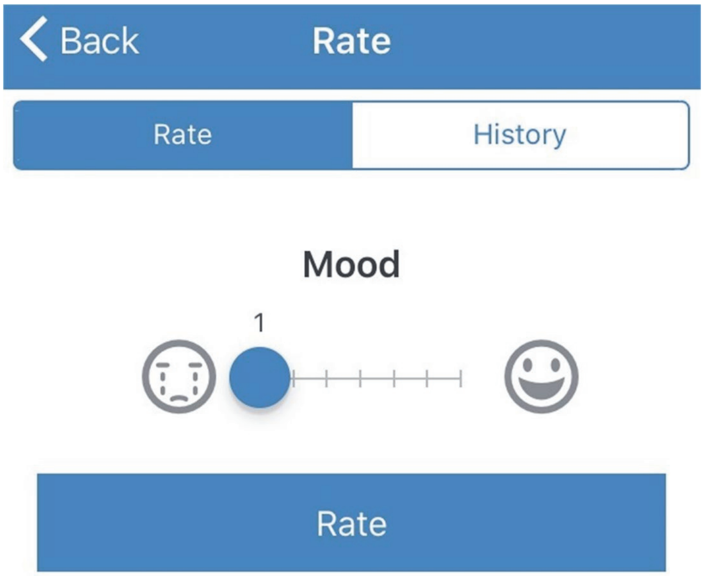

Figure 5. Screenshot of the Moodbuster smartphone application

2.9.8 Reasons for non-adherence

At the end of the study, at T1, participants will be asked online whether they completed the intervention and used it for the full duration of the study. If their response is negative, they will be asked to provide a rationale for not having completed the intervention or the study.

2.9.9 Level of engagement with the intervention

The third lesson is designed to stimulate users to schedule, perform, and evaluate pleasant activities. The number of these activities over time is assessed through log file analysis. Whether participants keep scheduling and recording activities for the duration of the study is an indicator of their engagement with the course, and of whether it has managed to make them more active.

2.9.10 Other measures

Screening for mental health issues will be performed before group allocation (T–1) using the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001), in order to deter people with more severe issues from taking part in the study. The PHQ-9 is composed of nine statements, each scored on a scale of 0 (not at all) to 3 (almost every day). Total scores range from 0 to 27, with higher scores indicating more severe depression, and scores over 14 moderate to severe depression (see Exclusion Criteria above). The PHQ-9 is considered to have good psychometric properties (Wittkampf et al., 2007).

2.10 Sample size

Since this study is a first in its sort, we know of no literature that indicates what effect size could be expected. Following the recommendation of Teare and colleagues (Teare et al., 2014), we plan to recruit 70 participants to determine the group means and standard deviations required for an estimation of the effect and sample sizes in a future RCT.

2.11 Statistical analysis

2.11.1 Primary analysis

The primary analysis will focus on the preliminary effectiveness of the virtual agent with respect to intervention adherence, as assessed in terms of intervention completion and mood recording response rates. Intervention completion will be assessed by calculating point estimates with corresponding 95% confidence intervals for both the intervention and the control group; a general linear model will be used to estimate the preliminary effect at the alpha < 0.05 significance level. That information will enable us to calculate the sample size required for observing a similar intervention effect in a larger RCT. To assess the mood recording response rate, we will conduct a logistic mixed-effects analysis to determine variations in adherence over time, following a similar analysis we performed in a previous ecological momentary assessment study (Provoost et al., 2018).

2.11.2 Secondary analysis

All secondary study parameters will be assessed with descriptive analysis, with formal tests merely serving to gain an estimation of possible group differences. Group differences will all be represented by point estimates and 95% confidence intervals. Within-group changes (pre–post, T0–T1) in motivation for taking part in the intervention (on the SMFL) and in mental health status (HADS-D) will also be tested formally with a mixed-effects model to estimate a time × group interaction effect and individual differences. Additionally, usability (SUS) and user satisfaction (CSQ-I) scores will be compared with the established benchmarks. Mood as measured by the smartphone records, and scheduled and recorded activities as measured in the platform, will only be analyzed descriptively. No subgroup analyses will be performed.

2.11.3 Data management

On the informed consent form, participants will be asked if they agree to the use of their data in future research on the same topic at VU University, and to their data being shared with regulatory authorities when required. This trial does not involve the collection of biological specimens for storage. All raw data will be stored on a secure local server at the VU University in Amsterdam, which is regularly backed up. Paper-based documents will be stored in a keycard-secured archive at the Department of Clinical, Neuro- and Developmental Psychology. All participants will be de-identified upon randomization by linking their participant number to a random study participant code. In the study, participants will be referred to exclusively by that participant code, and the document linking the two numbers will be destroyed once the study is over and results have been disseminated. Because this study is relatively small and investigator-initiated, no data monitoring committee or auditing process is required. Because we do not expect serious negative outcomes for the participants, we do not conduct an interim analysis and there are no subsequent formal stopping rules.

3 Discussion

The study described in this protocol paper is a pilot randomized controlled trial that will compare an unguided intervention for low mood with the same intervention with additional automated guidance provided in the form of a virtual coach. The main goal is to gain an estimate of the effectiveness of the virtual coach in terms of improving adherence to the intervention. That will help determine the feasibility and necessity of a future larger-scale trial.

Many studies have shown that online interventions that include human guidance are generally more effective than ones that do not. Human therapists or coaches that can provide such guidance are not always available, however, and the time of trained therapists is especially costly. Existing rules and protocols about providing guidance can be programmed into the interventions themselves so as to be automatically safeguarded. Moreover, automated support through ECAs enables human support factors such as empathy to be delivered more effectively. Automated support could improve adherence rates of guided, and especially of unguided, web-based interventions, and could thus improve their effectiveness.

While ECAs have been shown in many studies to be a feasible and acceptable technology in the domain of clinical psychology, very few applications have so far moved beyond the piloting phase. That is also the case for ECAs in iCBT contexts, where studies up to now have either been underpowered, have lacked control groups that set apart the ECA as the active ingredient, or have lacked depth in terms of underlying ECA technology. This study addresses these gaps in the literature in the following ways: (1) we designed a virtual coach that delivers automated support to iCBT for low mood; (2) we embedded it in an existing platform so that the platform can be used either with or without the ECA; and (3) we will estimate the effectiveness of a virtual coach in improving adherence and determine the parameters required for a proper RCT sample size calculation. Despite the technical limitations that come with embedding an ECA in an existing intervention platform, our virtual coach satisfies the criteria for an ECA – graphical embodiment, communicating with the user, and applying a form of reasoning – and conforms to recommendations from the literature. As a result, this study could represent a significant step in bridging the gap between guided and unguided iCBT interventions.

Acknowledgements

The authors would like to express their sincere thanks to Ms. Marleen Swenne for her help with designing the virtual coach, and to Mr. Ward van Breda for his help with the sentiment analysis algorithm.

CHAPTER 8

8

General Discussion

1 Aims of this thesis

The primary aim of this thesis was to determine whether, and how, an embodied conversational agent (ECA) could contribute to increasing adherence to unguided internet-based cognitive behavioral therapy (iCBT) for depression. Under the assumption that adherence is a mediating variable between human support and intervention effectiveness, the goal was to bridge the gap between guided and unguided iCBT for depression. This final chapter summarizes and integrates the main findings of the studies described in the previous chapters:

-What are the main findings of this thesis?

-What are the strengths and limitations of this thesis?

-What are the implications of the results?

-How do the results inform future research?

The chapter ends with a general conclusion of this thesis.

2 Main Findings

Below follows a brief summary of the main findings of this thesis. A more detailed description can be found in the individual chapters, while a more integrated summary is presented in the general conclusion at the end of this chapter.

2.1 Scoping Review

A scoping review on the use of ECAs in the treatment of common mental health disorders (Chapter 2) revealed an emerging field of research with most applications still in the explorative feasibility and piloting phases. Although results of these studies were generally positive in terms of system feasibility and user acceptance, evidence for the clinical effectiveness of ECAs in terms of improved adherence or clinical effects remained scarce for all disorders under consideration. No studies evaluated clinical outcomes in a setting that involved iCBT support provided by an ECA. As a way to increase the evidence base for ECAs in iCBT, we advocated a relatively 'low-tech' approach with more simplistic and less experimental ECA technology (e.g., cartoon-like instead of 3D animated embodiment, or menu-based instead of natural language user input) that could be rapidly developed, integrated with existing interventions, and tested.

2.2 EMA Adherence

The results of this experimental study, described in chapter 3, showed that the inclusion of a low-tech ECA providing feedback to ecological momentary assessment

(EMA) of mood did not improve adherence. Interestingly, however, while participants in the group without ECA feedback showed a lot of fluctuations in their adherence over time, adherence was much more stable in the ECA feedback condition. Given the nature of this study the stabilizing effect has to be interpreted with caution, but the paradigm that was applied, i.e., ECA feedback on EMA of mood, and mixed effects logistic analysis of adherence over time, can easily be re-applied to studies with different kinds of ECA feedback. In view of our recommendation from Chapter 2, this solution may have been too 'low-tech'. In an iCBT support context it seemed that a more sophisticated ECA would be more appropriate, and that the inclusion of this type of ECA feedback, specifically targeting EMA adherence, should not have the highest priority in the design of our own iCBT ECA solution.

2.3 Sentiment Analysis

In chapter 4, I compared sentiment analysis of open-text iCBT user input by people with major depression in a blended treatment by (1) a computer algorithm, with (2) human judgment by N=52 psychology students. The study showed that there was moderate agreement among the algorithm and human judges in terms of valence, and low agreement regarding specific emotions. Somewhat surprisingly, agreement among human judges showed very similar results, i.e., moderate agreement for valence and low agreement for specific emotions. This meant that the algorithm performed about as well as a randomly selected human judge when it came to evaluating overall sentiment of iCBT user input, and that it was less well suited for evaluating specific emotions. Taking human judgment as a golden standard, this indicated that automated analysis of overall sentiment could be considered in the development of an ECA for iCBT support.

2.4 Therapist Feedback

Chapter 5 described the results of a study into therapist feedback in a blended iCBT intervention for people with major depression. The therapist behaviors that occurred most frequently were informing (e.g., about the next session), encouraging (e.g., praising behavior), and affirming (e.g., normalizing behavior). Things therapists rarely did were making self-disclosures, confronting patients, and emphasizing patient responsibility. Therapists generally followed the feedback treatment protocol quite well, for example, by using correct greetings and endings of messages, scoring high on writing style, and limiting their feedback to 2 subjects. Regarding iCBT ECA support, this study showed which types of feedback were actually provided by therapists, and it also gave a clear account of the language they used to approach their clients. The protocol and examples could therefore be used to model and simulate automated feedback by an ECA that would resemble actual therapist feedback.

2.5 Therapist Adherence

While chapter 5 looked at therapist adherence to feedback protocols in a blended treatment for depression, with the focus on feedback content, chapter 6 aimed to establish therapist adherence to a blended treatment protocol for general and social anxiety as well as panic disorders. The main finding was that therapist adherence, or fidelity, in this study was high. This was exemplified by a ratio of face-to-face and online sessions (including feedback), and referencing to the online and face-to-face components, both as intended by the protocol. This study therefore provided more evidence that supportive humans, in this case licensed therapists, adhere well to iCBT feedback protocols. As the feedback protocols were a good reflection of what human support actually entailed, the results of this study provided us with more confidence in using feedback protocols as a basis upon which automated ECA support resembling human support could be modeled.

2.6 Moodbuster Lite

Chapter 7 described the protocol of a pilot randomized controlled trial (RCT) to evaluate Moodbuster Lite, a four-week iCBT intervention with integrated ECA support for people with low mood. The pilot study with N=70 participants will compare (1) Moodbuster Lite with ECA support with (2) Moodbuster Lite without ECA support, and its primary aim is to determine the effect of ECA support on adherence. As I mentioned in Chapter 1, the results of this study will be presented outside the scope of this thesis. In their absence, I will take this opportunity to briefly elaborate on the development process and the trajectory to ethical approval. The former is illustrative of how the first research goal, set in Chapter 1, of developing an ECA for iCBT was accomplished, while the latter is important to describe how the second research goal of assessing the supportive ECA in an iCBT context was not fully achieved within the time-span of this PhD project.

2.6.1 Development

Because this project explored a new area of research, it did not start out like most PhD projects in clinical psychology, where an intervention that is to be tested is usually known beforehand. Such projects typically involve different phases, like user testing through feasibility and pilot studies, possible adaptation of the intervention based on focus group research with end-users, and a final large RCT with sufficient statistical power to detect any effects of interest. Exploring a new domain, the first two years of this project went into defining how to shape the final phase, which meant we no longer had a full four years available to develop and test an intervention in the traditional way. Moreover, with the time left, defining the most optimal requirements for an ECA-solution by exploring the many different design choices that could be made through separate small studies, and then finding a suitable developer for the ECA, would have most likely resulted in what we observed all too often in our literature review: a genuinely impressive piece of technology that would require another PhD project to be integrated with an existing treatment platform and properly validated. It is for this reason that we made a set of pragmatic choices.

The goal of this project was to investigate whether a supportive ECA could improve adherence, and consequently effectiveness, of an iCBT intervention. With no available time to develop and test a new intervention, we looked for one that was (1) already validated, and (2) would allow for the integration of a new experimental component. Satisfying both criteria was the Moodbuster research and treatment platform, the result of a long-time collaboration with, among others, developers from the Portuguese Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC). The platform had already successfully been used to run a blended CBT (bCBT) for major depression intervention in the European E-COMPARED project, in which the effectiveness and cost-effectiveness of bCBT compared to treatment as usual in routine care were established (Kleiboer et al., 2016). We partnered up with the developers at INESC-TEC who were willing to help us integrate an ECA into the platform. Because we wanted to lower the threshold and time required for conducting a study, we started out by adjusting the Moodbuster for depression intervention to the context of low mood. With low mood being one of the primary symptoms of depression, this meant that the treatment mechanisms in the intervention could largely stay the same. The main adjustments were a significant reduction of the length of the intervention from a maximum of ten modules to three modules, and the adjustment of the language used to the context of low mood.

Next, knowing that ECAs were acceptable and feasible in clinical psychology interventions (Chapter 2), we put together what we could assume would work, and would be feasible to implement and evaluate in the remaining time of this project. Following insights from the general ECA literature and our own studies we designed a virtual coach that:

(1) could be easily embedded in the Moodbuster platform's webpages, similar to Youtube videos.
(2) could display emotions through facial expressions such that a sense of empathy could be simulated.
(3) could interact with users through dialogues based on motivational interviewing and iCBT feedback protocols (Chapters 5 and 6).

(4) could communicate with a sentiment analysis web-service (Chapter 4), such that open-text user input could be processed.
(5) could personalize conversations through user models based on previous interactions and user input to the platform.
(6) had its behavior represented in dialogue trees, such that it could be reviewed by domain experts.
(7) all in all, was a lot more sophisticated than the ECA from Chapter 3, which did not seem the most promising in an iCBT support context.

The final result of this development process was Moodbuster Lite: a well-validated formerly existing iCBT treatment platform, extended with an integrated and optional virtual coach that combines ECA technology, sentiment analysis, and extensive automated feedback, all of which represent a significant contribution to the iCBT for depression literature. Figure 1 below gives an overview of the final system and the interaction between its various components.

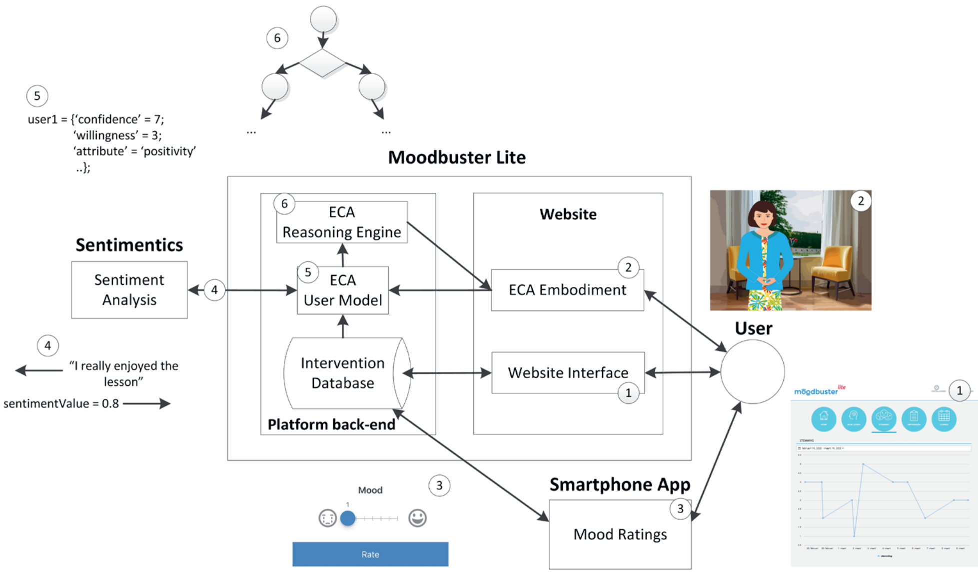

Figure 1. Overview of the Moodbuster Lite architecture

2.6.2 Ethical Approval

Having spent a month at the research institute of our Portuguese partners to integrate the virtual coach with the Moodbuster platform, we finished the intervention over the next couple of months to the extent that it was ready for the ethical approval process. Given the nature of our study design (pilot study), study population (people with low mood, but not depression), and intervention (short, low-intensity), we expected the study to be classified as not being subject to the laws and regulations for clinical studies involving a medical device, as this had been the case for a number of similar studies in our department that had recently received ethical approval. Around the time we were preparing our proposal, however, a number of changes were taking place. For example, the General Data Protection Regulation (GDPR) had taken effect around this time, and as a result, ethical committees had to become increasingly critical when it came to evaluating studies with online interventions. Under these new developments, Moodbuster Lite was classified as a medical device, and the study had to fulfill the full requirements for clinical research involving medical devices. In the end, the process of obtaining ethical approval ended up lasting 20 months, and involved well over 15 internal and external stakeholders, such as the medical ethical committee, a privacy champion, data protection officers, supervisors, an IT-consultant, domain experts, a translator, and experts in medical technology. Ethical approval was at last obtained, and the pilot RCT is expected to start in January 2021.

3 Strengths and Limitations

The main strength of this thesis lies in its interdisciplinary nature, combining clinical psychology with artificial intelligence (or more broadly computer science) research. Chapter 2 represented a first literature review with a systematic approach (scoping review) of ECAs in any context, thereby introducing a methodology traditionally used in the health sciences to this field of research, and the results of the studies in this thesis were published in scientific literature ranging from AI conference proceedings to psychology journals. Moreover, Moodbuster Lite, the final result of this thesis, integrates the two fields of research by embedding an ECA in an existing iCBT platform, with emphasis on both ECA technology and clinical effectiveness.

Another strength of this work is that the ECA that is integrated in Moodbuster Lite functions as an adjunct that can be turned on or off. This allows for a comparison that, to the best of our knowledge, has not yet been made involving an ECA with this level of sophistication: an intervention with ECA support versus the same intervention without ECA support. Making this comparison is important, as it allows us to control for any other variables, and thereby study the added benefits of ECA support in isolation. Both in applications where the ECA is the intervention (e.g., (Devault et al., 2014)), and where the ECA and intervention are integrated to the extent that it is not possible to deliver the intervention without it anymore (e.g., (Puskar et al., 2011; Burton et al., 2012)), such comparisons are not possible.

Limitations of the individual studies are discussed in the respective chapters, but one is relevant to this thesis as a whole. In shaping this project, we drew heavily on the conclusions from the literature review in Chapter 2. Being the first systematic review in this field, this made a lot of sense at the time, but progression in ECA research is fast-paced, and the search we conducted already dates back to July 2015. This raises the question whether our approach to Moodbuster Lite in Chapter 7 still holds up against the present evidence base, and what has happened in the field since then. One noticeable development that has taken place is that there has been an increased interest in chatbot applications (e.g., (Bendig et al., 2019)). These applications are often inspired by a Rogerian-dialogue approach through open-ended conversations in written natural language, as in the famous (very) early instantiation of this type called ELIZA (Weizenbaum, 1966). Chatbots have come a long way since then, and now apply advanced natural language processing, machine learning techniques, and therapeutic processes that are based on CBT. Although validation is still in its early phases, and more evidence regarding these unconstrained natural language input systems is needed (Laranjo et al., 2018; Abd-alrazaq et al., 2019; Bendig et al., 2019), these approaches have thus far seen some initial promising results in terms of depressive symptom reduction (Fitzpatrick et al., 2017; Fulmer et al., 2018; Inkster et al., 2018).

There have also been a number of more recent literature review papers on ECAs in health-care related domains that are closely related to our work. Two of them emphasize the importance of personalization, and highlight the sparsity of evidence regarding state-of-the-art approaches (Scholten et al., 2017), as well as the lack of theoretical underpinnings of personalization mechanisms (Kocaballi et al., 2019). We also stressed the importance of personalization through user modeling in Chapter 2, and made an effort to support the virtual coach's personalized feedback mechanisms both empirically (e.g., Chapters 4, 5 and 6) and with other literature. A review by Gaffney and colleagues (Gaffney et al., 2019) concluded that conversational agents are promising, but that more robust experimental designs (such as the one we apply in Chapter 7) are needed. Similar to Chapter 2, another review (Stal et al., 2020) concluded that there is little consensus over the optimal design features for ECAs in healthcare, and that although design features that promote the way in which ECAs are perceived seem positively related to usability and intention to use, there is little evidence for their effect on behavioral outcomes. A last review by Kramer et al. (Kramer et al., 2020) recommended a more systematic description of design activities in general, as well as a human-centered, stakeholder inclusive design approach. While we involved therapists and the IT-team responsible for the Moodbuster platform right from the start, and described the design of Moodbuster Lite's virtual coach in the protocol paper (Chapter 7), our pragmatic approach allowed little room for co-design with end-users. With the little time that was available, and with the intervention as well as the role (coach) and functionality (feedback and motivational interviewing protocols) of the ECA already established, we chose to focus our efforts on developing a working system based on the literature, and allow for plenty of flexibility based on user input later on.

4 Implications

Specifically, the studies in this thesis each come with their own implications:

Chapter 2. The literature review revealed a lack of evidence on the clinical effectiveness of ECAs used in the domain of clinical psychology, meaning that ECAs as they are applied presently are not yet ready for routine clinical practice, and that more rigorous studies are needed in this regard. Re-using solutions that are acceptable and feasible can prevent researchers from re-inventing the wheel.

Chapter 3. Although the results found in this study were ambiguous, studies using this kind of paradigm, i.e., investigating the effect of minimal (adjustments of) ECA technology on clinically relevant outcome measures such as adherence over time, can be helpful in further refining applications of ECAs in iCBT.

Chapter 4. Automated sentiment analysis of Dutch user input can be considered in iCBT applications, as the algorithm under study was about as accurate as a randomly selected human judge regarding overall valence. The disagreement among human judges regarding specific emotions contained in the texts can help inform a discussion about what appropriate benchmarks for computer algorithms should be in similar contexts.

Chapters 5 and 6. Therapists adhere well to iCBT feedback protocols, meaning that these kinds of protocols can be a good template to model human support for automated ECA delivery.

Chapter 7. It is feasible to design an ECA for iCBT support based on earlier research and guidelines, to develop it and integrate it with an existing treatment platform, and to study its effect on clinically relevant outcome measures in a randomized controlled study paradigm. Moodbuster Lite opens a path for studies on different ECA design aspects that can be guided from a technical (e.g., new machine learning feature), clinical (e.g., other ways of providing support), or user (e.g., personalized coach embodiment) perspective.

From a theoretical point of view, looking back at Figure 1 in Chapter 1, this thesis revolved around combining elements from human and technological support in ECA support (Figure 2). Chapter 2 revealed that most research up to that point had focused on the feasibility of designing ECAs for clinical psychology, how users related to them, and whether users would interact with them. In Chapter 3 we aimed to move beyond this by studying the relationship between ECA support and adherence. Chapters 4, 5, and 6 supported the transition from human support to ECA support, investigating whether and how elements from human iCBT support (sentiment recognition to give empathic responses, and the use of feedback protocols) could be applied in ECA support. The study proposed in Chapter 7 applies the lessons learned, and enables the exploration of the relationship between ECA support and adherence, as well as other variables such as the user-ECA relationship, engagement, motivation, intervention exposure, and effectiveness.

Figure 2. The hypothetical model from Chapter 1 revisited. This thesis combined elements from human and technological support in ECA support, aiming to improve the effectiveness of iCBT for depression by increasing adherence.

More generally, the aim of this thesis was to build a bridge between guided and unguided iCBT interventions for mood disorders by using ECAs. In practice, this also involved the building of another bridge, namely between computer scientists and clinical psychologists. The scoping review in chapter 2, by opening up the domain of ECAs to clinical psychologists (or vice versa), aimed to do precisely this, and the similar reviews of ECAs or related technologies in health-care settings that have been published since show that in this respect, this PhD project does not stand on its own. In the narrower context of ECAs in iCBT, and in view of the results of our review in Chapter 2 and the subsequent development process of Moodbuster Lite, some observations can be made that can serve as important considerations for both researchers and clinicians that are interested in building either of the aforementioned bridges with the use of ECA technology.

First of all, the focus in this thesis was on working with an existing intervention. The treatment mechanism, website-based iCBT, whether it is guided or unguided, already has a significant history of refinement and validation. The ECA is added to it as an adjunct, in order to improve something that we already know to be effective. Similar to how unguided interventions with smaller effect sizes than guided ones can be useful to society as a whole due to their larger uptake, effects of an ECA designed for unguided iCBT do not necessarily have to be large for the ECA to constitute a significant improvement. As is the case for Moodbuster Lite, when people do not like the ECA, they can simply ignore it, and are still left with the original effective intervention. Used in a real-world scenario, the intervention could also be offered without the virtual coach. As I discussed in chapter 2, the other approach to using ECAs in iCBT is by building an intervention around an ECA. It is the interaction with the ECA that constitutes the supposed treatment mechanism. 'Supposed', because as was shown in Chapter 2 and subsequent reviews described in Section 3 of this chapter, there is still little evidence for these types of interventions compared to website-based iCBT. Building up an evidence base takes a lot of time, and with no large RCTs to date that compared, for example, traditional face-to-face iCBT to 'ECA-based' iCBT, it seems fair to assume that uptake of these kinds of therapies in routine clinical practice is still far off. At present, therefore, clinical practice seems best served by ECA solutions that support, rather than deliver treatments.

Another point that I would like to make here is related to the technological sophistication of ECAs. ECAs can be more or less technologically advanced in either of their three dimensions: how they look (e.g., static pictures to 3D humanlike animations), how they communicate (e.g., text-based menus to spoken language), and how their behavior comes about (e.g., dialogue trees to machine learning). The most sophisticated ECAs use the most advanced technology, are able to do novel things that speak to the imagination, and are more likely to make their way into the media and shape people's expectations about ECAs (Gonzalez, 2017; Earley, 2020). When considering more sophisticated ECAs for website-based iCBT interventions, however, we run into a number of issues: (1) integration with existing intervention platforms becomes more difficult the more sophisticated an ECA is, since existing interventions were typically not designed with ECAs in mind; (2) for a variety of reasons (e.g., privacy in a household or bad quality of audio-visual input) communication by audio or video can be problematic for users in a real-world scenario; (3) a lot of sophisticated technologies such as text-processing operate by complex algorithms, resulting in a black box that cannot be interpreted by experts in the clinical domain for a safety check, (4) realistic graphics cannot always be processed by web-browsers (e.g., slow internet connections or computers), and may require third-party software resulting in long-term risks (e.g., Flashplayer which loses browser support by the end of 2020, used in (Swartout et al., 2013)), and (5) sophisticated ECAs require a lot more expertise overall, such as in graphics design, machine learning algorithms, or audio-visual processing, which can make long-term maintenance or even small adjustments difficult and costly to achieve. Therefore, currently, clinical practice seems best served by ECA solutions that take a more 'low-tech' approach and forego these kinds of issues.

Essentially, what I argue for, in order to increase the evidence base for ECAs in iCBT, speed up the research cycles, and bring ECA technology to routine clinical practice, is taking smaller steps. Smaller steps in terms of (1) what we expect ECAs to do (support instead of deliver), by improving what is already out there instead of re-inventing the wheel, and (2) in terms of ECA technology (low-tech instead of advanced), such that black box behavior can be avoided, patient safety can be guaranteed more easily, and solutions are applicable in real-world settings. Moodbuster Lite is my attempt to make this approach work, and shows that at the very least it is feasible to design, implement, and ethically justify a system in this manner. Even if the pilot RCT or a potential future RCT shows only a small positive effect of the ECA on adherence for a small group of participants, there is little reason not to consider it for application in routine clinical practice as long as no unforeseen negative effects occur. The approach can be considered equally well for other disorders for which there are both guided and unguided iCBT (or other) interventions, and allows for plenty of modifications if this is what future research requires.

I should emphasize that my recommendation for low-tech, and adjunctive ECA technology relates to the context of this thesis, i.e., the website-based iCBT as it is currently used to treat people with disorders such as depression. It is a recommendation for researchers and clinicians who want to apply ECAs in routine clinical practice right now, or in the foreseeable future. ECAs that aim to deliver iCBT such as Woebot (Fitzpatrick et al., 2017), or deliver elements that could be part of iCBT such as diagnosis by Simsensei Kiosk (Rizzo et al., 2016) are already being shown to be feasible with promising early results in terms of initial symptom reduction and diagnosis. With more and more budget being allocated to AI research, be it by big tech companies or national initiatives such as the Dutch AI Coalition, it seems a matter of time until more sophisticated ECAs become well validated or more easily accessible online such that they can become part of iCBT in routine clinical practice.

5 Future Research

The first and most obvious piece of future research has already been set into motion by conducting the pilot RCT, the goal and nature of which have been discussed at length. In the most optimal scenario both versions of Moodbuster Lite (with and without ECA support) will show some initial efficacy in terms of mood improvement, and adherence for the ECA support condition will be estimated to be slightly higher. Depending on the estimated effect of the ECA support, we can then determine the optimal sample size for a larger RCT and a more thorough validation involving a clinical population and outcome measure. Another interesting next step that involves the current intervention design, and that could be considered in a future RCT, is to compare the intervention with ECA support not only to the intervention without support, but also to the intervention with human support. Sufficiently powered RCTs like these could also help study the hypothetical model depicted in Figure 2 on a more detailed level by considering relevant variables such as therapeutic alliance, engagement and motivation, and compare the effects of various types of support, i.e., unguided, human, automated, or ECA support.

Besides a future large RCT, Moodbuster Lite offers various alternative paths for future research. As was described previously, we made many pragmatic design choices that are based on the literature, but have not been verified by including users in the design phase. We can expect the pilot RCT to inform some refinements to the design, but the current setup could also be adjusted to explore various design features of the ECA through experimental studies. Far from being the only ones, some examples are (1) the embodiment of the ECA, where we could experiment with different embodiments, genders, and cultural backgrounds, either by default or by user choice, (2) the ECA communication modalities, for example, by adding an optional spoken language feature, (3) different types of language use, possibly tailored towards user characteristics or user preferences, (4) further exploring the use of sentiment analysis beyond classifying feedback on intervention modules, (5) exploring therapeutic processes employed by the ECA through in- or excluding them in different versions, or (6) reducing or increasing the number of interactions with the virtual coach to find the optimal frequency.

More generally, there is an abundance of possible research trajectories involving ECAs and iCBT. I already touched upon the application of chatbots, which are rapidly gaining in interest, and at present are mostly used as stand-alone tools to deliver CBT. Although open-ended conversations are not well-suited for the delivery of intervention module-specific support, it is not hard to imagine chatbots as components of a more traditional website-based iCBT intervention. For example, a chatbot could be integrated in a website as an additional tool that is always available for more general CBT-related support aside from the actual intervention, and machine learning may be used to classify user input and match it with the most suitable answer similar to how chatbots for customer service are used (Suta et al., 2020). Once the technology that the general population has at its disposal improves and is widely available (e.g., VR-glasses, high-resolution webcams, or fast internet connections), state-of-the-art approaches that emphasize realism and believability could also become more relevant, either as components of website-based iCBT (e.g., intermediate automated assessments), or as a way to upgrade less realistic ECAs such as chatbots or Moodbuster Lite's virtual coach. A recent PhD thesis explored predictive modeling of future user states in iCBT (Van Breda, 2020), such as future mood or treatment success, and may well be integrated with an ECA to experiment with personalized timing of support. Lastly, integrating more advanced forms of user input in web-based iCBT also opens the door for other machine learning approaches, for example, sentiment analysis algorithms based on audio or video (Poria et al., 2016).

Given the somewhat limited resources at the disposal of psychologists, for example, in terms of IT-developers, an alternative approach to increasing the evidence base for ECAs in iCBT may be to seek more collaboration with partners outside of academia. Although it may not always be financially feasible to have them develop solutions from scratch, companies sometimes also develop software for mental health on their own accord that can greatly benefit from scientific input and validation. Such collaborations have the advantage that software development companies are usually experts in designing state-of-the-art systems, and are well-equipped to deal with constantly changing technologies. I already touched upon two examples of interventions that seem to have followed a similar path (Fitzpatrick et al., 2017; Fulmer et al., 2018), and it is promising to see that these kinds of collaborations are being stimulated with European grants. This is the case in, for example, the INTERREG eMen (eMen, 2020), or IT4Anxiety (IT4Anxiety, 2020) projects where small-to-medium sized companies receive support from mental health organizations and academia to develop, test, and implement e-mental health interventions.

6 Conclusion

I started out this thesis asking whether an ECA that simulates human support factors can make iCBT more effective. A literature review (Chapter 2) was conducted to explore this new area of research, and it was concluded that at present, ECA applications seemed feasible and acceptable, that evidence regarding their effectiveness was sparse, and that there were no examples where an ECA supported iCBT users in a manner similar to how a human would do so. Based on this early exploration I defined two research goals: (1) to design an ECA based on what is known to work, and (2) to evaluate whether such an ECA could contribute to increasing iCBT adherence. To reach the first goal I conducted and participated in a number of studies, which showed that a supportive ECA for iCBT would require at least some level of sophistication (Chapter 3), that a sentiment analysis algorithm could be used to classify the valence of open-text user input (Chapter 4), and that therapist protocols for providing iCBT support could be applied in designing ECA dialogues (Chapters 5 and 6). Together with insights from the ECA literature and input from domain experts, these findings were used to design a virtual coach that was integrated with an existing iCBT intervention platform. A pilot RCT study involving Moodbuster Lite, the resulting intervention, was designed (Chapter 7) to accomplish the second research goal. While I argue that state-of-the-art ECA approaches are not yet ready for application in routine clinical practice, this thesis shows that there is plenty of less advanced ECA technology that is safer to use, easier to develop and maintain, and can be considered in an iCBT context. Future research with Moodbuster Lite first and foremost involves the pilot RCT, which may inform a future large RCT. The platform also offers plenty of opportunities to conduct more research on ECA design features. As AI and technological developments are fast-paced, more advanced approaches using, for example, unconstrained natural language in- and output or highly realistic embodiments, are likely to start playing a larger role in iCBT in the near future. To finish with the metaphor used throughout this thesis, the research presented shows how a bridge between guided and unguided iCBT interventions for depression was built in the form of Moodbuster Lite. Future improvements to the bridge will be informed by the results of the pilot RCT.

CHAPTER 9

9

Summary

1 Introduction

Guided internet-based cognitive therapy (iCBT) for depression is considered more effective than unguided iCBT, which may be explained by the higher adherence rates found for the guided format. Unguided iCBT, on the other hand, is potentially more scalable and accessible as it does not require human involvement. This thesis explores whether this gap between unguided and guided interventions, i.e., effectiveness versus scalability and accessibility, can be bridged using embodied conversational agents (ECAs). ECAs are more or less autonomous virtual characters with an embodiment used to communicate with a user, and may be used to provide automated iCBT support. With endless possible variations in ECA embodiments, ways of communicating, and reasoning capabilities, there is no clear solution to designing an ECA for iCBT support. The two research goals of this thesis, elaborated upon in **Chapter 1**, are to (1) design and develop an ECA for iCBT support based on what is known to work and (2) establish whether such an ECA can increase adherence to unguided iCBT.

2 Embodied conversational agents in clinical psychology

As a first step, **Chapter 2** presents the results of a scoping review into the use of ECAs in the treatment of common mental health disorders. We reviewed N=49 studies that presented primary research findings on ECA applications that targeted mood, anxiety, psychotic, autism spectrum, or substance use disorders. More than half of the studies (n=26) focused on autism treatment, and ECAs were used most often for social skills training (n=23). Most applications (n=43) were still in the development and piloting phases, and few studies conducted controlled research into clinical effects of ECAs, ECAs used in the context of iCBT, or ECAs used as an adjunct to existing interventions. To increase the evidence base with regard to ECAs for iCBT support, we proposed an additional focus on low-tech ECA solutions that can be rapidly developed, tested, and applied in routine practice.

3 Experimental studies

We followed up on the literature review by conducting two experimental studies to further inform the design of our ECA. In the first of these, described in **Chapter 3**, we investigated whether feedback from a minimalistic ECA could increase adherence. In a three-week ecological momentary assessment smartphone study, N=68 participants were asked to report their mood three times a day. An ECA could mirror participant-reported mood states. A two-stage randomization into a text and ECA feedback group, versus a text-only control group, was applied to control for individual differences and feedback history. While the ECA's feedback did not increase adherence, adherence in the intervention group remained more constant than in the control group. Although this was a pilot study with a non-clinical population, and the results should therefore be interpreted with some caution, the study shows how ECA feedback may have a stabilizing effect on adherence, how controlled experiments on the relationship between ECA support and clinically relevant measures such as adherence may be conducted, and how adherence over time can be analyzed.

In the second experimental study, described in **Chapter 4**, we validated a sentiment analysis algorithm for the Dutch language against human judgment in an iCBT context. Automated sentiment analysis could allow an ECA to give tailored feedback to free text user input, but it was unclear how well the technology would be applicable to iCBT. N=493 iCBT user texts were evaluated on overall sentiment and the presence of five specific emotions by the algorithm, and by N=52 human judges who evaluated 75 randomly selected texts each. Inter-rater agreement between the algorithm and human judges was found to be moderate with respect to overall sentiment (ICC = 0.55, $p < .01$), and low with respect to the specific emotions. Interestingly, agreement among human judges showed a similar pattern, i.e., moderate for overall sentiment (ICC = 0.58, $p < .01$), and low for the specific emotions. This meant that the algorithm performed about as well as a randomly selected human judge. Taking average human judgment as a benchmark, automated analysis of overall sentiment could therefore be considered for practical application.

4 Feedback in iCBT support

In the next two studies we looked more closely into the content of real-world iCBT feedback to inform the design of our ECA dialogues. **Chapter 5** presents a study that aimed to identify which therapist behaviors occur in written online feedback, assess how well therapists adhere to feedback instructions, and explore whether either is associated with patient outcomes. In this study we analyzed a total of N=219 therapist feedback messages for patients in a blended CBT (bCBT) for depression intervention. The most frequently used therapist behaviors were informing, encouraging, and affirming, but no clear relationships between behaviors and therapy outcomes were found. Two feedback aspects, structure and readability, were negatively correlated with session completion ($\rho = -.34$, $p = .02$ and $\rho = -.36$, $p = .02$, respectively). Therapists adhered well to most feedback instructions, which meant they could be used as a template to model real-world iCBT feedback for our ECA dialogues.

Chapter 6 presents another study into the content of therapist feedback and adherence to feedback instructions. Specifically, it aimed to assess therapist fidelity to the principles of a bCBT protocol for anxiety disorders. N=257 therapist feedback messages and transcripts of N=74 face-to-face sessions were analyzed to determine the ratio of face-to-face to online sessions, therapy intensity, and therapist adherence to bCBT feedback instructions. Correlations were found between patients' share of online sessions and both treatment intensity (r = .37, p < .05), as well as patient computer experience (r = .31, p < .05). Overall therapist fidelity was high, providing more evidence for the applicability of iCBT feedback instructions in modeling realistic ECA dialogues.

5 Moodbuster Lite

The research from the previous chapters comes together in the development of Moodbuster Lite, a 4-week iCBT intervention for low mood with integrated ECA support. **Chapter 7** describes the protocol of a pilot two-armed randomized controlled trial that evaluates the intervention. The study will assess the effectiveness of virtual coach support in terms of improved intervention adherence, as well as the feasibility of a future, larger-scale trial with Moodbuster Lite. Secondary aims will be to assess the virtual coach's effect on motivation, user perceptions of the virtual coach, and the general feasibility of the intervention. N=70 participants from the general population who wish to learn how they can improve their mood will be recruited and randomized to either (1) Moodbuster Lite with ECA support, or (2) Moodbuster Lite without ECA support. Candidates with symptoms of moderate to severe depression will be excluded from study participation. Assessments will be taken at baseline and post-study, after four weeks.

6 Discussion

Chapter 8 provides a general discussion of the research presented in this thesis, including the main findings, strengths, limitations, and implications, as well as suggestions for future research. From the perspective of this thesis as a whole, the research goals were to design an ECA for iCBT based on what is known to work, and to evaluate whether such an ECA could contribute to increasing iCBT adherence. The design of the ECA that was developed to function as a virtual coach in iCBT, was largely informed by the research conducted in Chapters 2 to 6; it has a reasonable amount of technological sophistication (Chapter 3), applies automated sentiment analysis (Chapter 4), interacts according to iCBT feedback protocols (Chapters 5 and 6), and targets a clinically relevant outcome (Chapter 2), i.e., adherence. After the subsequent integration of the ECA with the Moodbuster Lite intervention, and having obtained formal ethical approval, Moodbuster Lite is now ready for evaluation in a pilot randomized controlled trial as described in Chapter 7. The intervention and study design show how a bridge between guided and unguided iCBT interventions was built in the form of Moodbuster Lite, and the pilot RCT will inform how the bridge can be further improved.

References

References

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* 132, 103978. doi:10.1016/j.ijmedinf.2019.103978.

Agarwal, R., Sampath, H. A., and Indurkhya, B. (2013). "A Usability Study on Natural Interaction Devices with ASD Children," in *Proceedings of the 7th International Conference on Universal Access in Human-Computer Interaction: User and Context Diversity - Volume 2* UAHCI'13., eds. C. Stephanidis and M. Antona (Berlin, Heidelberg: Springer-Verlag), 447–453. doi:10.1007/978-3-642-39191-0_49.

Aggarwal, R., and Singh, H. (2013). Differential Influence of Blogs Across Different Stages of Decision Making: The Case of Venture Capitalists. *MIS Q.* 37, 1093–1112. doi:10.25300/MISQ/2013/37.4.05.

Alcorn, A., Pain, H., Rajendran, G., Smith, T., Lemon, O., Porayska-Pomsta, K., et al. (2011). "Social Communication between Virtual Characters and Children with Autism," in *Artificial Intelligence in Education* Lecture Notes in Computer Science., eds. G. Biswas, S. Bull, J. Kay, and A. Mitrovic (Springer Berlin Heidelberg), 7–14. doi:10.1007/978-3-642-21869-9_4.

Alemi, M., Meghdari, A., Ghanbarzadeh, A., Moghadam, L. J., and Ghanbarzadeh, A. (2014). "Impact of a Social Humanoid Robot as a Therapy Assistant in Children Cancer Treatment," in *Social Robotics SE - 2* Lecture Notes in Computer Science., eds. M. Beetz, B. Johnston, and M.-A. Williams (Springer International Publishing), 11–22. doi:10.1007/978-3-319-11973-1_2.

Amethier, P., and Haggren, H. (2018). Data-Driven Predictions of Outcome for an Internet-Delivered Treatment Against Anxiety Disorders: A Comparison of Clinician and Algorithm Performance. Available at: https://kth.diva-portal.org/smash/get/diva2:1218565/FULLTEXT01.pdf.

Amirabdollahian, F., Robins, B., Dautenhahn, K., and Ji, Z. (2011). "Investigating tactile event recognition in child-robot interaction for use in autism therapy.," in *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* (United States: IEEE), 5347–51. doi:10.1109/IEMBS.2011.6091323.

An, L. C., Demers, M. R. S., Kirch, M. A., Considine-Dunn, S., Nair, V., Dasgupta, K., et al. (2013). A randomized trial of an avatar-hosted multiple behavior change intervention for young adult smokers. *J. Natl. Cancer Inst. Monogr.* 2013, 209–15.

doi:10.1093/jncimonographs/lgt021.

Andersson, G., and Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cogn. Behav. Ther.* 38, 196–205. doi:10.1080/16506070903318960.

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., and Hedman, E. (2014). Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. *World Psychiatry* 13, 288–295. doi:10.1002/wps.20151.

Andersson, G., and Titov, N. (2014). Advantages and limitations of Internet-based interventions for common mental disorders. *World Psychiatry* 13, 4–11. doi:10.1002/wps.20083.

Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., et al. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *J. Anxiety Disord.* 55, 70–78. doi:10.1016/j.janxdis.2018.01.001.

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

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi:10.1080/1364557032000119616.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.* 34, 555–596. doi:10.1162/coli.07-034-R2.

Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., and Riper, H. (2016). Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study. *J. Med. Internet Res.* 18, e72. doi:10.2196/jmir.5505.

ATLAS.ti Scientific Software Development GmBH, Berlin, Germany Available at: https://atlasti.com/.

Bamasak, O., Braik, R., Al-Tayari, H., Al-Harbi, S., Al-Semairi, G., and Abu-Hnaidi, M. (2013). "Improving Autistic Children's Social Skills Using Virtual Reality," in *Design, User Experience, and Usability. Health, Learning, Playing, Cultural, and Cross-Cultural User Experience SE - 38* Lecture Notes in Computer Science., ed. A. Marcus (Springer Berlin Heidelberg), 342–351. doi:10.1007/978-3-642-39241-2_38.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *Int. J. Hum. Comput. Interact.* 24, 574–594. doi:10.1080/10447310802205776.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67, 1–48. doi:10.18637/jss.v067.i01.

Baylor, A. L. (2011). The design of motivational agents and avatars. *Educ. Technol. Res. Dev.* 59, 291–300. doi:10.1007/s11423-011-9196-3.

Baylor, A. L., and Kim, S. (2009). Designing nonverbal communication for pedagogical agents: When less is more. *Comput. Human Behav.* 25, 450–457. doi:10.1016/j.chb.2008.10.008.

Baylor, A. L., and Kim, Y. (2004). "Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role," in *Intelligent Tutoring Systems*, 592–603. doi:10.1007/978-3-540-30139-4_56.

Beck, A. T., Epstein, N., Brown, G., and Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *J. Consult. Clin. Psychol.* 56, 893–897. doi:10.1037/0022-006X.56.6.893.

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

Bekele, E., Crittendon, J. A., Swanson, A., Sarkar, N., and Warren, Z. E. (2014). Pilot clinical application of an adaptive robotic system for young children with autism. *Autism* 18, 598–608. doi:10.1177/1362361313479454.

Bellg, A. J., Resnick, B., Minicucci, D. S., Ogedegbe, G., Ernst, D., Borrelli, B., et al. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Heal. Psychol.* 23, 443–451. doi:10.1037/0278-6133.23.5.443.

Bendig, E., Erb, B., Schulze-Thuesing, L., and Baumeister, H. (2019). The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review. *Verhaltenstherapie*, 1–13. doi:10.1159/000501812.

Bernardini, S., Porayska-Pomsta, K., and Smith, T. J. (2014). ECHOES: An intelligent serious game for fostering social communication in children with autism. *Inf. Sci. (Ny).* 264, 41–60. doi:10.1016/j.ins.2013.10.027.

Beutler, L. E., Malik, M. L., Alimohamed, S., Harwood, T. M., Talebi, H., and Noble, S. (2004). "Therapist variables," in *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*, ed. M. J. Lambert (New York: Wiley), 227–306.

Bickmore, T., and Gruber, A. (2010). Relational Agents in Clinical Psychiatry. *Harv. Rev. Psychiatry* 18, 119–130. doi:10.3109/10673221003707538.

Bickmore, T., Schulman, D., and Shaw, G. (2009a). DTask and litebody: Open source, standards-based tools for building web-deployed embodied conversational agents. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 5773 LNAI, 425–431. doi:10.1007/978-3-642-04380-2_46.

Bickmore, T., Schulman, D., and Sidner, C. L. (2009b). Issues in Designing Agents for Long-Term Behavior Change. *CHI Work. Engagem. by Des.*, 1–5.

Bickmore, T., Schulman, D., and Yin, L. (2010a). Maintaining Engagement in Long-Term Interventions With Relational Agents. *Appl. Artif. Intell.* 24, 648–666. doi:10.1080/08839514.2010.492259.

Bickmore, T. W., Mitchell, S. E., Jack, B. W., Paasche-Orlow, M. K., Pfeifer, L. M., and Odonnell, J. (2010b). Response to a Relational Agent by Hospital Patients with Depressive Symptoms. *Interact. Comput.* 22, 289–298. doi:10.1016/j.intcom.2009.12.001.

Bickmore, T. W., and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Interact.* 12, 293–327. doi:10.1145/1067860.1067867.

Bickmore, T. W., Puskar, K., Schlenk, E. A., Pfeifer, L. M., and Sereika, S. M. (2010c). Maintaining reality: Relational agents for antipsychotic medication adherence. *Interact. Comput.* 22, 276–288. doi:10.1016/j.intcom.2010.02.001.

Boccanfuso, L., and O'Kane, J. M. (2010). Adaptive Robot Design with Hand and Face Tracking for Use in Autism Therapy. in *Social Robotics* Lecture Notes in Computer Science., eds. S. Ge, H. Li, J.-J. Cabibihan, and Y. Tan (Singapore: Springer Berlin Heidelberg), 265–274. doi:10.1007/978-3-642-17248-9_28.

Boß, L., Lehr, D., Reis, D., Vis, C., Riper, H., Berking, M., et al. (2016). Reliability and Validity of Assessing User Satisfaction With Web-Based Health Interventions. *J. Med. Internet Res.* 18, e234. doi:10.2196/jmir.5952.

Bosse, T., and Provoost, S. (2015). "Integrating Conversation Trees and Cognitive Models Within an ECA for Aggression De-escalation Training," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 650–659. doi:10.1007/978-3-319-25524-8_48.

Bosse, T., Van Breda, W., Van Dijk, N., and Scholte, J. (2017). An agent-based aggression de-escalation training application for football referees. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 10423 LNAI, 3–14. doi:10.1007/978-3-319-65340-2_1.

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., et al. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *J. Consult. Clin. Psychol.* 81, 443–454. doi:10.1037/a0031437.

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Int. J. Hum. Comput. Stud.* 62, 161–178. doi:http://dx.doi.org/10.1016/j.ijhcs.2004.11.002.

Bresó, A., Martínez-Miranda, J., Botella, C., Baños, R. M., and García-Gómez, J. M. (2016). Usability and acceptability assessment of an empathic virtual agent to prevent major depression. *Expert Syst.* 33, 297–312. doi:10.1111/exsy.12151.

Broderick, J. E., Schwartz, J. E., Shiffman, S., Hufford, M. R., and Stone, A. a (2003). Signaling does not adequately improve diary compliance. *Ann. Behav. Med.* 26, 139–48. doi:10.1207/S15324796ABM2602_06.

Broekens, J., Heerink, M., and Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology* 8, 94–103. doi:10.4017/gt.2009.08.02.002.00.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Eval. Ind.* 189, 4–7.

Bruinsma, A., Kampman, M., Exterkate, C. C., and Hendriks, G. J. (2016). An exploratory study of "blended" cognitive behavioural therapy (CBT) for patients with a panic disorder: results and patients' experiences. *Tijdschr. Psychiatr.* 58, 361–270.

Burger, F., Neerincx, M. A., and Brinkman, W.-P. (2020). Technological State of the Art of Electronic Mental Health Interventions for Major Depressive Disorder: Systematic Literature Review. *J. Med. Internet Res.* 22, e12599. doi:10.2196/12599.

Burleson, W. (2006). Affective Learning Companions: strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance. Available at: http://hdl.handle.net/1721.1/37404.

Burton, C., Wolters, M. K., Serrano Blanco, A., Szentagotai, A., Ure, J., Pagliari, C., et al. (2012). Help4Mood: avatar-based support for treating people with major depression in the community. *Int. J. Integr. Care* 12, e51. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571175/.

Canidate, S., and Hart, M. (2017). The Use of Avatar Counseling for HIV/AIDS Health Education: The Examination of Self-Identity in Avatar Preferences. *J. Med. Internet Res.* 19, e365. doi:10.2196/jmir.6740.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., and Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cogn. Behav. Ther.* 47, 1–18. doi:10.1080/16506073.2017.1401115.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., and Balain, S. (2007). A conceptual framework for implementation fidelity. *Implement. Sci.* 2, 1–9. doi:10.1186/1748-5908-2-40.

Carroll, K. M., and Kiluk, B. D. (2017). Cognitive behavioral interventions for alcohol and drug use disorders: Through the stage model and back again. *Psychol.*

Addict. Behav. 31, 847–861. doi:10.1037/adb0000311.

chatbots.org (2020). Available at: https://www.chatbots.org/synonyms/#all.

Cheek, C., Bridgman, H., Fleming, T., Cummings, E., Ellis, L., Lucassen, M. F., et al. (2014). Views of Young People in Rural Australia on SPARX, a Fantasy World Developed for New Zealand Youth With Depression. *JMIR serious games* 2, e3. doi:10.2196/games.3183.

Chen, J., Chen, D., Wang, L., and Lemon, O. (2010). "An Advanced Learning Environment Aided by Recognition of Multi-modal Social Signals," in *Advances in Web-Based Learning – ICWL 2010 SE* - 5 Lecture Notes in Computer Science., eds. X. Luo, M. Spaniol, L. Wang, Q. Li, W. Nejdl, and W. Zhang (Springer Berlin Heidelberg), 41–51. doi:10.1007/978-3-642-17407-0_5.

Christensen, H., Griffiths, K. M., and Farrer, L. (2009). Adherence in Internet Interventions for Anxiety and Depression. *J. Med. Internet Res.* 11, e13. doi:10.2196/jmir.1194.

Clavel, C., and Callejas, Z. (2016). Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. *IEEE Trans. Affect. Comput.* 7, 74–93. doi:10.1109/TAFFC.2015.2444846.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Jiyong Ma, Movellan, J., et al. (2003). Perceptive animated interfaces: first steps toward a new paradigm for human-computer interaction. *Proc. IEEE* 91, 1391–1405. doi:10.1109/JPROC.2003.817143.

Costa, S., Lehmann, H., Dautenhahn, K., Robins, B., and Soares, F. (2015). Using a Humanoid Robot to Elicit Body Awareness and Appropriate Physical Interaction in Children with Autism. *Int. J. Soc. Robot.* 7, 265–278. doi:10.1007/s12369-014-0250-2.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 337, a1655. doi:10.1136/bmj.a1655.

Cuijpers, P. (1997). Bibliotherapy in unipolar depression: A meta-analysis. *J. Behav. Ther. Exp. Psychiatry* 28, 139–147. doi:10.1016/S0005-7916(97)00005-0.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., and Dobson, K. S. (2013a). A Meta-Analysis of Cognitive-Behavioural Therapy for Adult Depression, Alone and in Comparison with other Treatments. *Can. J. Psychiatry* 58, 376–385. doi:10.1177/070674371305800702.

Cuijpers, P., and Dekker, J. (2005). Psychologische behandeling van depressie; een systematisch overzicht van meta-analysen. *Ned Tijdschr Geneeskd* 149, 1892–1897.

Cuijpers, P., Huibers, M., Daniel Ebert, D., Koole, S. L., and Andersson, G. (2013b).

How much psychotherapy is needed to treat depression? A metaregression analysis. *J. Affect. Disord.* 149, 1–13. doi:10.1016/j.jad.2013.02.030.

Cuijpers, P., Koole, S. L., van Dijke, A., Roca, M., Li, J., and Reynolds, C. F. (2014). Psychotherapy for subclinical depression: meta-analysis. *Br. J. Psychiatry* 205, 268–274. doi:10.1192/bjp.bp.113.138784.

Cuijpers, P., van Straten, A., Warmerdam, L., and Andersson, G. (2008). Psychological treatment of depression: a meta-analytic database of randomized studies. *BMC Psychiatry* 8, 36. doi:10.1186/1471-244X-8-36.

Daudt, H. M., van Mossel, C., and Scott, S. J. (2013). Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Med. Res. Methodol.* 13, 48. doi:10.1186/1471-2288-13-48.

Davis, K., Drey, N., and Gould, D. (2009). What are scoping studies? A review of the nursing literature. *Int. J. Nurs. Stud.* 46, 1386–400. doi:10.1016/j.ijnurstu.2009.02.010.

de Bruin, E. J., and Meijer, A. M. (2017). The impact of online therapeutic feedback on outcome measures in Internet-CBTI for adolescents with insomnia. *Sleep Med.* 29, 68–75. doi:10.1016/j.sleep.2016.05.017.

Desmet, P. M. A., Vastenburg, M. H., Van Bel, D., and Romero, N. (2012). Pick-a-Mood Development and Application of a Pictorial Mood-Reporting Instrument. *Proc. 8th Int. Des. Emot. Conf. London 2012 Cent. Saint Martins Coll. Art Des.*, 11–14.

Devault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. *2014 Int. Conf. Auton. Agents Multi-Agent Syst. Int. Found. Auton. Agents Multiagent Syst.*, 1061–1068. doi:10.1016/j.imavis.2005.08.005.

Devitt, A., and Ahmad, K. (2007). Sentiment polarity identification in financial news: a cohesion-based approach. *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 984–991. doi:10.1.1.143.7157.

Dickerson, P., Robins, B., and Dautenhahn, K. (2013). Where the action is: A conversation analytic perspective on interaction between a humanoid robot, a co-present adult and a child with an ASD. *Interact. Stud.* 14, 297–316. doi:10.1075/is.14.2.07dic.

Diehl, J. J., Schmitt, L. M., Villano, M., and Crowell, C. R. (2012). The Clinical Use of Robots for Individuals with Autism Spectrum Disorders: A Critical Review. *Res. Autism Spectr. Disord.* 6, 249–262. doi:10.1016/j.rasd.2011.05.006.

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., and Glozier, N. (2011). A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. *J. Med. Internet Res.* 13, e52. doi:10.2196/jmir.1772.

Earley, K. (2020). Humour is the best medicine for Woebot's mental health chatbot. *Siliconrepublic*. Available at: https://www.siliconrepublic.com/start-ups/woebot-alison-darcy-psychology-technology [Accessed September 21, 2020].

Efird, J. (2010). Blocked Randomization with Randomly Selected Block Sizes. *Int. J. Environ. Res. Public Health* 8, 15–20. doi:10.3390/ijerph8010015.

Ellis, J. D., Grekin, E. R., Beatty, J. R., McGoron, L., LaLiberte, B. V., Pop, D. E., et al. (2017). Effects of narrator empathy in a computer delivered brief intervention for alcohol use. *Contemp. Clin. Trials* 61, 29–32. doi:10.1016/j.cct.2017.07.014.

eMen (2020). e-mental health innovation and transnational implementation platform North West Europe (eMEN). Available at: https://www.nweurope.eu/projects/project-search/e-mental-health-innovation-and-transnational-implementation-platform-north-west-europe-emen/ [Accessed March 25, 2020].

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. . (2002). Structured Clinical Interview for DSM-IV-TR Axis I Disorders. *New York State Psychiatr. Inst.*

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Heal.* 4, e19. doi:10.2196/mental.7785.

Fujimoto, I., Matsumoto, T., De Silva, P. R. S., Kobayashi, M., and Higashi, M. (2011). Mimicking and Evaluating Human Motion to Improve the Imitation Skill of Children with Autism Through a Robot. *Int. J. Soc. Robot.* 3, 349–357. doi:10.1007/s12369-011-0116-9.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., and Rauws, M. (2018). Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment. Heal.* 5, e64. doi:10.2196/mental.9782.

Fydrich, T., Dowdall, D., and Chambless, D. L. (1992). Reliability and validity of the beck anxiety inventory. *J. Anxiety Disord.* 6, 55–61. doi:10.1016/0887-6185(92)90026-4.

Gaffney, H., Mansell, W., and Tai, S. (2019). Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Ment. Heal.* 6, e14166. doi:10.2196/14166.

Gamer, M., Lemon, J., and Fellows Puspendra Singh, I. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. Available at: https://cran.r-project.org/package=irr.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., and Ngeow, E.

(2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clin. Psychol. Rev.* 31, 79–88. doi:10.1016/j.cpr.2010.09.007.

Gega, L., Marks, I., and Mataix-Cols, D. (2004). Computer-Aided CBT Self-Help for Anxiety and Depressive Disorders: Experience of a London Clinic and Future Directions. *J. Clin. Psychol.* 60, 147–157. doi:10.1002/jclp.10241.

Gerritsen, C., and Van Breda, W. R. J. (2015). "Simulation-Based Prediction and Analysis of Collective Emotional States," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 118–126. doi:10.1007/978-3-319-20367-6_13.

Gonzalez, R. (2017). Virtual Therapists Help Veterans Open Up About PTSD. *WIRED, Sci.* Available at: https://www.wired.com/story/virtual-therapists-help-veterans-open-up-about-ptsd/ [Accessed September 21, 2020].

Google (2018). Google Cloud Natural Language. Available at: https://cloud.google.com/natural-language/ [Accessed June 29, 2018].

Gratch, J., and Hartholt, A. (2013). Virtual Humans: A New Toolkit for Cognitive Science Research. *Proc. Cogn. Sci. Soc.* 35, 41–42. Available at: http://escholarship.org/uc/item/9j9404c3.

Griffiths, K. M., and Christensen, H. (2007). Internet-based mental health programs: A powerful tool in the rural medical kit. *Aust. J. Rural Health* 15, 81–87. doi:10.1111/j.1440-1584.2007.00859.x.

Grolleman, J., van Dijk, B., Nijholt, A., and van Emst, A. (2006). "Break the Habit! Designing an e-Therapy Intervention Using a Virtual Coach in Aid of Smoking Cessation," in *Persuasive Technology* Lecture Notes in Computer Science., eds. W. IJsselsteijn, Y. DeKort, C. Midden, B. Eggen, and E. VandenHoven, 133–141. doi:10.1007/11755494_19.

Gun, S. Y., Titov, N., and Andrews, G. (2011). Acceptability of internet treatment of anxiety and depression. *Australas. Psychiatry* 19, 259–264. doi:10.3109/10398562.2011.562295.

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., et al. (2013). "All Together Now," in *Intelligent Virtual Agents SE - 33* Lecture Notes in Computer Science., eds. R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira (Springer Berlin Heidelberg), 368–381. doi:10.1007/978-3-642-40415-3_33.

Hatcher, R. L., and Gillaspy, J. A. (2006). Development and validation of a revised short version of the working alliance inventory. *Psychother. Res.* 16, 12–25. doi:10.1080/10503300500352500.

Hayes, A. F., and Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.* 1, 77–89.

doi:10.1080/19312450709336664.

Hill, C. E., Helms, J. E., Tichenor, V., Spiegel, S. B., O'Grady, K. E., and Perry, E. S. (1988). Effects of therapist response modes in brief psychotherapy. *J. Couns. Psychol.* 35, 222–233. doi:10.1037/0022-0167.35.3.222.

Holländare, F., Gustafsson, S. A., Berglind, M., Grape, F., Carlbring, P., Andersson, G., et al. (2016). Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. *Internet Interv.* 3, 1–7. doi:10.1016/j.invent.2015.11.002.

Hopkins, I. M., Gower, M. W., Perez, T. A., Smith, D. S., Amthor, F. R., Casey Wimsatt, F., et al. (2011). Avatar assistant: Improving social skills in students with an asd through a computer-based intervention. *J. Autism Dev. Disord.* 41, 1543–1555. doi:10.1007/s10803-011-1179-z.

Hsieh, H.-F., and Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qual. Health Res.* 15, 1277–1288. doi:10.1177/1049732305276687.

Hudlicka, E. (2013). Virtual training and coaching of health behavior: example from mindfulness meditation training. *Patient Educ. Couns.* 92, 160–166. doi:10.1016/j.pec.2013.05.007.

IBM (2018). IBM Watson Natural Language Understanding. Available at: https://www.ibm.com/watson/services/natural-language-understanding/ [Accessed June 29, 2018].

Inkster, B., Sarda, S., and Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth uHealth* 6, e12106. doi:10.2196/12106.

Isbister, K., and Doyle, P. (2004). "The Blind Men and the Elephant Revisited," in *From Brows to Trust: Evaluating Embodied Conversational Agents*, eds. Z. Ruttkay and C. Pelachaud (Dordrecht: Springer Netherlands), 3–26. doi:10.1007/1-4020-2730-3_1.

IT4Anxiety (2020). IT4Anxiety - Managing anxiety via innovative technologies for better mental health. Available at: https://www.nweurope.eu/projects/project-search/it4anxiety-managing-anxiety-via-innovative-technologies-for-better-mental-health/ [Accessed October 1, 2020].

Jochems, E. C. (2016). Motivation for Psychiatric Treatment in Outpatients with Severe Mental Illness Different Perspectives. doi:10.13140/RG.2.1.3496.2805.

Johansson, R., and Andersson, G. (2012). Internet-based psychological treatments for depression. *Expert Rev. Neurother.* 12, 861–870. doi:10.1586/ern.12.63.

Jordan, K., King, M., Hellersteth, S., Wiren, A., Mulligan, H., Wirén, A., et al. (2013).

Feasibility of using a humanoid robot for enhancing attention and social skills in adolescents with autism spectrum disorder. *Int. J. Rehabil. Res.* 36, 221–227. doi:10.1097/MRR.0b013e32835d0b43.

Josephine, K., Josefine, L., Philipp, D., David, E., and Harald, B. (2017). Internet- and mobile-based depression interventions for people with diagnosed depression: A systematic review and meta-analysis. *J. Affect. Disord.* 223, 28–40. doi:10.1016/j.jad.2017.07.021.

Kang, S.-H., and Gratch, J. (2010). Virtual humans elicit socially anxious interactants' verbal self-disclosure. *Comput. Animat. Virtual Worlds* 21, 473–482. doi:10.1002/cav.345.

Kang, S.-H., and Gratch, J. (2012). Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories. *Stud. Health Technol. Inform.* 181, 202–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22954856.

Karyotaki, E., Efthimiou, O., Miguel, C., Bermpohl, F. M. genannt, Furukawa, T. A., Cuijpers, P., et al. (2021). Internet-Based Cognitive Behavioral Therapy for Depression. *JAMA Psychiatry*. doi:10.1001/jamapsychiatry.2020.4364.

Karyotaki, E., Furukawa, T. A., Efthimiou, O., Riper, H., and Cuijpers, P. (2019). Guided or self-guided internet-based cognitive–behavioural therapy (iCBT) for depression? Study protocol of an individual participant data network meta-analysis. *BMJ Open* 9, e026820. doi:10.1136/bmjopen-2018-026820.

Karyotaki, E., Kleiboer, a., Smit, F., Turner, D. T., Pastor, a. M., Andersson, G., et al. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. *Psychol. Med.* 45, 2717–2726. doi:10.1017/S0033291715000665.

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., et al. (2017). Efficacy of Self-guided Internet-Based Cognitive Behavioral Therapy in the Treatment of Depressive Symptoms. *JAMA Psychiatry* 74, 351. doi:10.1001/jamapsychiatry.2017.0044.

Keijsers, G. P. J., Minnen, A. van, and Hoogduin, C. A. L. (2010). *Protocollaire behandelingen voor volwassenen met psychische klachten. Deel 1.*

Keijsers, G. P. J., Schaap, C. P. D. R., and Hoogduin, C. A. L. (2000). The Impact of Interpersonal Patient and Therapist Behavior on Outcome in Cognitive-Behavior Therapy. *Behav. Modif.* 24, 264–297. doi:10.1177/0145445500242006.

Kelders, S. M. (2015). "Involvement as a Working Mechanism for Persuasive Technology," in *Persuasive Technology SE - 1* Lecture Notes in Computer Science., eds. T. MacTavish and S. Basapur (Cham: Springer International Publishing), 3–14. doi:10.1007/978-3-319-20306-5_1.

Kelders, S. M., Kok, R. N., Ossebaard, H. C., and Van Gemert-Pijnen, J. E. (2012). Persuasive System Design Does Matter: a Systematic Review of Adherence to Web-based Interventions. *J. Med. Internet Res.* 14, e152. doi:10.2196/jmir.2104.

Kemmeren, L. L., Schaik, D. J. F. Van, Riper, H., Kleiboer, A. M., Bosmans, J. E., and Smit, J. H. (2016). Effectiveness of blended depression treatment for adults in specialised mental healthcare : study protocol for a randomised controlled trial. *BMC Psychiatry*, 1–14. doi:10.1186/s12888-016-0818-5.

Kemmeren, L. L., Van Schaik, D. J. F., Smit, J. H., Ruwaard, J., Rocha, A., Henriques, M. R., et al. (2019). Unraveling the black box: Exploring usage patterns of a blended treatment for depression in a multicenter study. *J. Med. Internet Res.* 21. doi:10.2196/12707.

Kenter, R. M. F., van de Ven, P. M., Cuijpers, P., Koole, G., Niamat, S., Gerrits, R. S., et al. (2015). Costs and effects of Internet cognitive behavioral treatment blended with face-to-face treatment: Results from a naturalistic study. *Internet Interv.* 2, 77–83. doi:10.1016/j.invent.2015.01.001.

Kleiboer, A., Donker, T., Seekles, W., van Straten, A., Riper, H., and Cuijpers, P. (2015). A randomized controlled trial on the role of support in Internet-based problem solving therapy for depression and anxiety. *Behav. Res. Ther.* 72, 63–71. doi:10.1016/j.brat.2015.06.013.

Kleiboer, A., Smit, J., Bosmans, J., Ruwaard, J., Andersson, G., Topooco, N., et al. (2016). European COMPARative Effectiveness research on blended Depression treatment versus treatment-as-usual (E-COMPARED): study protocol for a randomized controlled, non-inferiority trial in eight European countries. *Trials* 17, 387. doi:10.1186/s13063-016-1511-1.

Knox, S., and Hill, C. E. (2003). Therapist self-disclosure: Research-based suggestions for practitioners. *J. Clin. Psychol.* 59, 529–539. doi:10.1002/jclp.10157.

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., et al. (2019). The Personalization of Conversational Agents in Health Care: Systematic Review. *J. Med. Internet Res.* 21, e15360. doi:10.2196/15360.

Konstantinidis, E. I., Hitoglou-Antoniadou, M., Luneski, A., Bamidis, P. D., and Nikolaidou, M. M. (2009). Using affective avatars and rich multimedia content for education of children with autism. in *Proceedings of the 2nd International Conference on PErvsive Technologies Related to Assistive Environments - PETRA '09* PETRA '09. (New York, New York, USA: ACM Press), 1–6. doi:10.1145/1579114.1579172.

Kooistra, L. C., Wiersma, J. E., Ruwaard, J., van Oppen, P., Smit, F., Lokkerbol, J., et al. (2014). Blended vs. face-to-face cognitive behavioural treatment for major depression in specialized mental health care: study protocol of a randomized controlled cost-effectiveness trial. *BMC Psychiatry* 14, 290. doi:10.1186/s12888-014-0290-z.

Kramer, L. L., ter Stal, S., Mulder, B. C., de Vet, E., and van Velsen, L. (2020). Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review. *J. Med. Internet Res.* 22, e14058. doi:10.2196/14058.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9. *J. Gen. Intern. Med.* 16, 606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

Ku, J., Han, K., Lee, H. R., Jang, H. J., Kim, K. U., Park, S. H., et al. (2007). VR-Based Conversation Training Program for Patients with Schizophrenia: A Preliminary Clinical Trial. *CyberPsychology Behav.* 10, 567–574. doi:10.1089/cpb.2007.9989.

Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., and Sarkar, N. (2011). "Design of a Virtual Reality Based Adaptive Response Technology for Children with Autism Spectrum Disorder," in *Affective Computing and Intelligent Interaction SE - 20* Lecture Notes in Computer Science., eds. S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin (Springer Berlin Heidelberg), 165–174. doi:10.1007/978-3-642-24600-5_20.

Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 159. doi:10.2307/2529310.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., et al. (2018). Conversational agents in healthcare: a systematic review. *J. Am. Med. Informatics Assoc.* 25, 1248–1258. doi:10.1093/jamia/ocy072.

Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., and Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Eval. Program Plann.* 2, 197–207. doi:10.1016/0149-7189(79)90094-6.

Leff, J., Williams, G., Huckvale, M. A., Arbuthnot, M., and Leff, A. P. (2013). Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study. *Br. J. Psychiatry* 202, 428–433. doi:10.1192/bjp.bp.112.124883.

Levac, D., Colquhoun, H., and O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implement. Sci.* 5, 69. doi:10.1186/1748-5908-5-69.

Lewinsohn, P. M., Biglan, A., and Zeiss, A. M. (1976). "Behavioral treatment of depression," in *The behavioral management of anxiety, depression and pain*, ed. P. O. Davidson (New York: Brunner/Mazel), 91–146.

Lisetti, C., Amini, R., Yasavur, U., and Rishe, N. (2013). I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Trans. Manag. Inf. Syst.* 4, 1–28. doi:10.1145/2544103.

Lopez, M. A., and Basco, M. A. (2015). Effectiveness of Cognitive Behavioral Therapy in Public Mental Health: Comparison to Treatment as Usual for Treatment-

Resistant Depression. *Adm. Policy Ment. Heal. Ment. Heal. Serv. Res.* 42, 87–98. doi:10.1007/s10488-014-0546-4.

Martínez-Miranda, J., Bresó, A., and García-Gómez, J. M. (2014). "Look on the Bright Side: A Model of Cognitive Change in Virtual Agents," in *Intelligent Virtual Agents SE - 37* Lecture Notes in Computer Science., eds. T. Bickmore, S. Marsella, and C. Sidner (Springer International Publishing), 285–294. doi:10.1007/978-3-319-09767-1_37.

Mastermind (2018). Mastermind Homepage. Available at: http://mastermind-project.eu.

Mathers, C. D., and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* 3, 2011–2030. doi:10.1371/journal.pmed.0030442.

Mays, N., Roberts, E., and Popay, J. (2001). "Synthesising research evidence," in *Studying the organisation and delivery of health services: Research methods*, eds. N. Fulop, P. Allen, A. Clarke, and N. Black (London: Routledge), 188–220.

McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., and Otto, M. W. (2013). Patient Preference for Psychological vs Pharmacologic Treatment of Psychiatric Disorders. *J. Clin. Psychiatry* 74, 595–602. doi:10.4088/JCP.12r07757.

McKeachie, W. J., and Svinicki, M. D. (2010). *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers (College Teaching Series)*. Belmont: Wadsworth.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* 5, 1093–1113. doi:10.1016/j.asej.2014.04.011.

Merry, S. N., Stasiak, K., Shepherd, M., Frampton, C., Fleming, T., and Lucassen, M. F. G. (2012). The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. *BMJ* 344, e2598. doi:10.1136/bmj.e2598.

Milne, M., Powers, D., and Leibbrandt, R. (2009). Development of a software-based social tutor for children with autism spectrum disorders. in *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group on Design: Open 24/7 - OZCHI '09* OZCHI '09. (New York, New York, USA: ACM Press), 265–268. doi:10.1145/1738826.1738870.

Minddistrict, Amsterdam, The Netherlands Available at: https://www.minddistrict.com/.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, c869–c869.

doi:10.1136/bmj.c869.

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

Mohr, D. C., Hart, S. L., Howard, I., Julian, L., Vella, L., Catledge, C., et al. (2006). Barriers to psychotherapy among depressed and nondepressed primary care patients. *Ann. Behav. Med.* 32, 254–258. doi:10.1207/s15324796abm3203_12.

Mol, M., Dozeman, E., Provoost, S., van Schaik, A., Riper, H., and Smit, J. H. (2018). Behind the Scenes of Online Therapeutic Feedback in Blended Therapy for Depression: Mixed-Methods Observational Study. *J. Med. Internet Res.* 20, e174. doi:10.2196/jmir.9890.

Mol, M., Dozeman, E., van Schaik, D. J. F., Vis, C. P. C. D., Riper, H., and Smit, J. H. (2016). The therapist's role in the implementation of internet-based cognitive behavioural therapy for patients with depression: study protocol. *BMC Psychiatry* 16, 338. doi:10.1186/s12888-016-1045-9.

Mol, M., van Genugten, C., Dozeman, E., van Schaik, D. J. F., Draisma, S., Riper, H., et al. (2019). Why Uptake of Blended Internet-Based Interventions for Depression Is Challenging: A Qualitative Study on Therapists' Perspectives. *J. Clin. Med.* 9, 91. doi:10.3390/jcm9010091.

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

Morie, J. F., Antonisse, J., Bouchard, S., and Chance, E. (2009). Virtual worlds as a healing modality for returning soldiers and Veterans. *Annu. Rev. CyberTherapy Telemed.* 7, 273–276. doi:10.3233/978-1-60750-017-9-273.

movisensXS, Version 0.7.4162, Karlsruhe, Germany Available at: https://xs.movisens.com/.

Mumm, J., and Mutlu, B. (2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Comput. Human Behav.* 27, 1643–1650. doi:10.1016/j.chb.2011.02.002.

Myhr, G., and Payne, K. (2006). Cost-Effectiveness of Cognitive-Behavioural Therapy for Mental Disorders: Implications for Public Health Care Funding Policy in Canada. *Can. J. Psychiatry* 51, 662–670. doi:10.1177/070674370605101006.

Nakao, S., Nakagawa, A., Oguchi, Y., Mitsuda, D., Kato, N., Nakagawa, Y., et al. (2018). Web-based cognitive behavioral therapy blended with face-to-face sessions for major depression: Randomized controlled trial. *J. Med. Internet Res.* 20, 1–15. doi:10.2196/10743.

Nass, C. I., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. *Comput. Interact. Conf. Celebr. Interdepend. 1994*, 72–78. doi:10.1145/259963.260288.

National Institute for Health and Clinical Excellence. (2011). Generalised anxiety disorder and panic disorder in adults: Management. Clinical guideline 113. Availabe from: www.nice.org.uk/guidance/cg113.

Oinas-Kukkonen, H., and Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Commun. Assoc. Inf. Syst.* 24, 485–500.

Owen, J., and Hilsenroth, M. J. (2014). Treatment adherence: The importance of therapist flexibility in relation to therapy outcomes. *J. Couns. Psychol.* 61, 280–288. doi:10.1037/a0035753.

Pagliari, C., Burton, C., McKinstry, B., Szentatotai, A., David, D., Serrano Blanco, A., et al. (2012). Psychosocial implications of avatar use in supporting therapy for depression. *Stud. Health Technol. Inform.* 181, 329–333. doi:10.3233/978-1-61499-121-2-329.

Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in Virtual Agents and Robots. *ACM Trans. Interact. Intell. Syst.* 7, 1–40. doi:10.1145/2912150.

Palestra, G., Bortone, I., Cazzato, D., Adamo, F., Argentiero, A., Agnello, N., et al. (2014). "Social Robots in Postural Education: A New Approach to Address Body Consciousness in ASD Children," in *Social Robotics SE - 30* Lecture Notes in Computer Science., eds. M. Beetz, B. Johnston, and M.-A. Williams (Springer International Publishing), 290–299. doi:10.1007/978-3-319-11973-1_30.

Pang, B., and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends® Inf. Retr.* 2, 1–135. doi:10.1561/1500000011.

Paxling, B., Lundgren, S., Norman, A., Almlöv, J., Carlbring, P., Cuijpers, P., et al. (2013). Therapist Behaviours in Internet-Delivered Cognitive Behaviour Therapy: Analyses of E-Mail Correspondence in the Treatment of Generalized Anxiety Disorder. *Behav. Cogn. Psychother.* 41, 280–289. doi:10.1017/S1352465812000240.

Persons, Jacqueline B.; Bertagnolli, A. (1999). Inter-Rater Reliability of Cognitive - Behavioral Case Formulations of Depression: A Replication. *Cognit. Ther. Res.* 23, 271–283. doi:10.1023/A:1018791531158.

Persons, J. B., Mooney, K. A., and Padesky, C. A. (1995). Interrater relability of cognitive-behavioral case formulations. *Cognit. Ther. Res.* 19, 21–34. doi:10.1007/BF02229674.

Pinto, M. D., Greenblatt, A. M., Hickman, R. L., Rice, H. M., Thomas, T. L., and Clochesy, J. M. (2016). Assessing the Critical Parameters of eSMART-MH: A Promising Avatar-Based Digital Therapeutic Intervention to Reduce Depressive Symptoms. *Perspect. Psychiatr. Care* 52, 157–68. doi:10.1111/ppc.12112.

Pontier, M., and Siddiqui, G. F. (2008). "A Virtual Therapist That Responds Empathically to Your Answers," in *Intelligent Virtual Agents* IVA '08. (Berlin, Heidelberg: Springer Berlin Heidelberg), 417–425. doi:10.1007/978-3-540-85483-8_42.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., and Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174, 50–59. doi:10.1016/j.neucom.2015.01.095.

Portney, L., and Watkins, M. (2000). *Foundations of Clinical Research: Applications to Practice*. Michigan: Prentice Hall Health.

Provoost, S., Kleiboer, A., Ornelas, J., Bosse, T., Ruwaard, J., Rocha, A., et al. (2020). Improving adherence to an online intervention for low mood with a virtual coach: study protocol of a pilot randomized controlled trial. *Trials* 21, 860. doi:10.1186/s13063-020-04777-2.

Provoost, S., Lau, H. M., Ruwaard, J., and Riper, H. (2017). Embodied Conversational Agents in Clinical Psychology: A Scoping Review. *J. Med. Internet Res.* 19, e151. doi:10.2196/jmir.6553.

Provoost, S., Ruwaard, J., Neijenhuijs, K., Bosse, T., and Riper, H. (2018). "Mood Mirroring with an Embodied Virtual Agent: A Pilot Study on the Relationship Between Personalized Visual Feedback and Adherence," in *Communications in Computer and Information Science*, 24–35. doi:10.1007/978-3-319-94779-2_3.

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., and Bosse, T. (2019). Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts: An Exploratory Study. *Front. Psychol.* 10, 1–12. doi:10.3389/fpsyg.2019.01065.

Prowse, P. T. (2015). A Meta-Evaluation: The Role of Treatment Fidelity within Psychosocial Interventions during the Last Decade. *J. Psychiatry* 18. doi:10.4172/Psychiatry.1000251.

Puskar, K., Schlenk, E. A., Callan, J., Bickmore, T., and Sereika, S. (2011). Relational agents as an adjunct in schizophrenia treatment. *J. Psychosoc. Nurs. Ment. Health Serv.* 49, 22–9. doi:10.3928/02793695-20110705-01.

R Core Team (2016). R: A Language and Environment for Statistical Computing. Available at: https://www.r-project.org.

Raghavan, V. V., and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *J. Am. Soc. Inf. Sci.* 37, 279–287. doi:10.1002/(SICI)1097-4571(198609)37:5<279::AID-ASI1>3.0.CO;2-Q.

Reeves, B., and Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York, NY, US: Cambridge University Press.

Renaud, J., Russell, J. J., and Myhr, G. (2014). Predicting Who Benefits Most From Cognitive-Behavioral Therapy for Anxiety and Depression. *J. Clin. Psychol.* 70, 924–932. doi:10.1002/jclp.22099.

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. Available at: https://cran.r-project.org/package=psych.

Ribeiro, P. C., Araujo, B. B. P. L. de, and Raposo, A. (2014). ComFiM: A Cooperative Serious Game to Encourage the Development of Communicative Skills Between Children with Autism. in *Proceedings of the 2014 Brazilian Symposium on Computer Games and Digital Entertainment* SBGAMES '14. (Washington, DC, USA: IEEE Computer Society), 148–157. doi:10.1109/SBGAMES.2014.19.

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

Rinck, M., Rörtgen, T., Lange, W.-G., Dotsch, R., Wigboldus, D. H. J., and Becker, E. S. (2010). Social anxiety predicts avoidance behaviour in virtual encounters. *Cogn. Emot.* 24, 1269–1276. doi:10.1080/02699930903309268.

Ring, L., Bickmore, T., and Pedrelli, P. (2016). An affectively aware virtual therapist for depression counseling. *ACM SIGCHI Conf. Hum. Factors Comput. Syst. Work. Comput. Ment. Heal.*

Riper, H., Andersson, G., Christensen, H., Cuijpers, P., Lange, A., and Eysenbach, G. (2010). Theme Issue on E-Mental Health: A Growing Field in Internet Research. *J. Med. Internet Res.* 12, e74. doi:10.2196/jmir.1713.

Rizzo, A., Lucas, G., Gratch, J., Stratou, G., Morency, L.-P., Chavez, K., et al. (2016). Automatic Behavior Analysis During a Clinical Interview with a Virtual Human. *Stud. Health Technol. Inform.* 220, 316–322. doi:10.3233/978-1-61499-625-5-316.

Robins, B., and Dautenhahn, K. (2014). Tactile Interactions with a Humanoid Robot: Novel Play Scenario Implementations with Children with Autism. *Int. J. Soc. Robot.* 6, 397–415. doi:10.1007/s12369-014-0228-0.

Rollnick, S., Miller, W. R., and Butler, C. C. (2008). *Motivational Interviewing in Health Care: Helping Patients Change Behavior*. New York: The Guilford Press.

Romijn, G., Batelaan, N., Kok, R., Koning, J., van Balkom, A., Titov, N., et al. (2019). Internet-Delivered Cognitive Behavioral Therapy for Anxiety Disorders in Open Community Versus Clinical Service Recruitment: Meta-Analysis. *J. Med. Internet Res.* 21, e11706. doi:10.2196/11706.

Romijn, G., Batelaan, N., Koning, J., van Balkom, A., de Leeuw, A., Benning, F., et al. Acceptability, effectiveness and cost-effectiveness of blended cognitive-behavioural therapy (bCBT) for anxiety disorders in specialised mental health care: A 15-week randomised controlled trial with 1-year follow-up (under review).

Romijn, G., Riper, H., Kok, R., Donker, T., Goorden, M., van Roijen, L. H., et al. (2015). Cost-effectiveness of blended vs. face-to-face cognitive behavioural therapy for severe anxiety disorders: study protocol of a randomized controlled trial. *BMC Psychiatry* 15, 311. doi:10.1186/s12888-015-0697-1.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., et al. (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. *Biol. Psychiatry* 54, 573–583. doi:10.1016/S0006-3223(02)01866-8.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi:10.1037/h0077714.

Ruttkay, Z., Dormann, C., and Noot, H. (2004). "Embodied Conversational Agents on a Common Ground," in *From Brows to Trust: Evaluating Embodied Conversational Agents* Human-Computer Interaction Series., eds. Z. Ruttkay and C. Pelachaud (Springer Netherlands), 27–66. doi:10.1007/1-4020-2730-3_2.

Ruttkay, Z., and Pelachaud, C. (2004). *From Brows To Trust: Evaluating Embodied Conversational Agents*. Dordrecht: Kluwer Academic Publishers.

Ruwaard, J., Schrieken, B., Schrijver, M., Broeksteeg, J., Dekker, J., Vermeulen, H., et al. (2009). Standardized Web-Based Cognitive Behavioural Therapy of Mild to Moderate Depression: A Randomized Controlled Trial with a Long-Term Follow-Up. *Cogn. Behav. Ther.* 38, 206–221. doi:10.1080/16506070802408086.

Ryan, R. M., and Deci, E. L. (2008). A self-determination theory approach to psychotherapy: The motivational basis for effective change. *Can. Psychol. Can.* 49, 186–193. doi:10.1037/a0012753.

Sánchez-Ortiz, V. C., Munro, C., Startup, H., Treasure, J., and Schmidt, U. (2011). The Role of Email Guidance in Internet-Based Cognitive-Behavioural Self-Care Treatment for Bulimia Nervosa. *Eur. Eat. Disord. Rev.* 19, 342–348. doi:10.1002/erv.1074.

Schmidt, R., Eifler, P., and Masuch, M. (2013). "Playfully Conquering Performance Anxiety," in *Games for Health*, eds. B. Schouten, S. Fedtke, T. Bekker, M. Schijven, and A. Gekker (Wiesbaden: Springer Fachmedien Wiesbaden), 267–279. doi:10.1007/978-3-658-02897-8_21.

Schneider, L. H., Hadjistavropoulos, H. D., and Faller, Y. N. (2016). Internet-delivered Cognitive Behaviour Therapy for Depressive Symptoms: An Exploratory Examination of Therapist Behaviours and their Relationship to Outcome and Therapeutic Alliance. *Behav. Cogn. Psychother.* 44, 625–639. doi:10.1017/S1352465816000254.

Scholten, M. R., Kelders, S. M., and Van Gemert-Pijnen, J. E. (2017). Self-Guided Web-

Based Interventions: Scoping Review on User Needs and the Potential of Embodied Conversational Agents to Address Them. *J. Med. Internet Res.* 19, e383. doi:10.2196/jmir.7351.

Schueller, S. M., Tomasino, K. N., and Mohr, D. C. (2017). Integrating Human Support Into Behavioral Intervention Technologies: The Efficiency Model of Support. *Clin. Psychol. Sci. Pract.* 24, 27–45. doi:10.1111/cpsp.12173.

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, c332–c332. doi:10.1136/bmj.c332.

Sentimentics, Rotterdam, The Netherlands Available at: http://www.sentimentics.com.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. in *Journal of Clinical Psychiatry*, 22–33. doi:10.1016/S0924-9338(99)80239-9.

Shiffman, S., Stone, A. a, and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415.

Shoukry, L., Caserman, P., Göbel, S., and Steinmetz, R. (2015). “Blitzmerker: Learning Idioms with a Mobile Game,” in *Serious Games SE - 7* Lecture Notes in Computer Science., eds. S. Göbel, M. Ma, J. Baalsrud Hauge, M. F. Oliveira, J. Wiemeyer, and V. Wendel (Springer International Publishing), 77–88. doi:10.1007/978-3-319-19126-3_7.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability.1. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979, 86:420–8. *Psychol. Bull.* 86, 420–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18839484.

Smith, C., Cavazza, M., Charlton, D., Zhang, L., Turunen, M., and Hakulinen, J. (2008). “Integrating Planning and Dialogue in a Lifestyle Agent,” in *Intelligent Virtual Agents* Lecture Notes in Computer Science., eds. H. Prendinger, J. Lester, and M. Ishizuka (Berlin, Heidelberg: Springer Berlin Heidelberg), 146–153. doi:10.1007/978-3-540-85483-8_15.

Smith, M. J., Ginger, E. J., Wright, K., Wright, M. A., Taylor, J. L., Humm, L. B., et al. (2014a). Virtual reality job interview training in adults with autism spectrum disorder. *J. Autism Dev. Disord.* 44, 2450–63. doi:10.1007/s10803-014-2113-y.

Smith, M. J., Ginger, E. J., Wright, M., Wright, K., Boteler Humm, L., Olsen, D., et al. (2014b). Virtual reality job interview training for individuals with psychiatric disabilities. *J. Nerv. Ment. Dis.* 202, 659–67. doi:10.1097/NMD.0000000000000187.

Snaith, R. P., and Zigmond, A. S. (1986). The hospital anxiety and depression scale. *BMJ* 292, 344–344. doi:10.1136/bmj.292.6516.344.

Sobkowicz, P., Kaschesky, M., and Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Gov. Inf. Q.* 29, 470–479. doi:10.1016/j.giq.2012.06.005.

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

Spijker, J., Bockting, C., Meeuwissen, J. A., Van Vliet, I. M., Emmelkamp, P. M., and Hermens, M. L. (2013). *Multidisciplinaire richtlijn Depressie (Derde revisie): Richtlijn voor de diagnostiek, behandeling en begeleiding van volwassen patiënten met een depressieve stoornis*. Utrecht: Trimbos Institute.

Spinhoven, P., Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., and Hemert, A. M. VAN (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychol. Med.* 27, S0033291796004382. doi:10.1017/S0033291796004382.

SPSS Inc., Chicago IL, USA Available at: https://www.ibm.com/analytics/spss-statistics-software.

Stal, S. ter, Kramer, L. L., Tabak, M., Akker, H. op den, and Hermens, dr. ir. H. (2020). Design Features of Embodied Conversational Agents in eHealth: a Literature Review. *Int. J. Hum. Comput. Stud.*, 102409. doi:10.1016/j.ijhcs.2020.102409.

Stein, M. B., and Stein, D. J. (2008). Social anxiety disorder. *Lancet* 371, 1115–1125. doi:10.1016/S0140-6736(08)60488-2.

Stewart, R. E., and Chambless, D. L. (2009). Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: a meta-analysis of effectiveness studies. *J. Consult. Clin. Psychol.* 77, 595–606. doi:10.1037/a0016032.

Stinckens, N., Ulburghs, A., and Claes, L. (2009). De werkalliantie als sleutelelement in het therapiegebeuren. *Tijdschr voor Klin Psychol.* 39, 44–60. Available at: https://www.han.nl/onderzoek/werkveld/projecten/alliantiegericht-werken-v/werkwijze/_attachments/2009_-_stinckens_ea_-_de_werkalliantie_als_sleuteleleement_in_het_therapiegebeuren_1_.pdf.

Suganuma, S., Sakamoto, D., and Shimoyama, H. (2018). An Embodied Conversational Agent for Unguided Internet-Based Cognitive Behavior Therapy in Preventative Mental Health: Feasibility and Acceptability Pilot Trial. *JMIR Ment. Heal.* 5, e10454. doi:10.2196/10454.

Suta, P., Lan, X., Wu, B., Mongkolnam, P., and Chan, J. H. (2020). An overview of machine learning in chatbots. *Int. J. Mech. Eng. Robot. Res.* 9, 502–510. doi:10.18178/ijmerr.9.4.502-510.

Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., et al. (2013). Virtual Humans for Learning. *Ai Mag.* 34, 13–30. doi:10.1609/aimag.v34i4.2487.

Sztein, D. M., Koransky, C. E., Fegan, L., and Himelhoch, S. (2018). Efficacy of cognitive behavioural therapy delivered over the Internet for depressive symptoms: A systematic review and meta-analysis. *J. Telemed. Telecare* 24, 527–539. doi:10.1177/1357633X17717402.

Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., et al. (2015). Automated Social Skills Trainer. in *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15* IUI '15. (New York, New York, USA: ACM Press), 17–27. doi:10.1145/2678025.2701368.

Teare, M., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., and Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 15, 264. doi:10.1186/1745-6215-15-264.

Thase, M. E., Wright, J. H., Eells, T. D., Barrett, M. S., Wisniewski, S. R., Balasubramani, G. K., et al. (2018). Improving the Efficiency of Psychotherapy for Depression: Computer-Assisted Versus Standard CBT. *Am. J. Psychiatry* 175, 242–250. doi:10.1176/appi.ajp.2017.17010089.

Tielman, M., Brinkman, W., and Neerincx, M. A. (2014). "Design Guidelines for a Virtual Coach for Post-Traumatic Stress Disorder Patients," in *Intelligent Virtual Agents*, 434–437. doi:10.1007/978-3-319-09767-1_54.

Titov, N., Andrews, G., Davies, M., McIntyre, K., Robinson, E., and Solley, K. (2010). Internet Treatment for Depression: A Randomized Controlled Trial Comparing Clinician vs. Technician Assistance. *PLoS One* 5, e10939. doi:10.1371/journal.pone.0010939.

Titzler, I., Saruhanjan, K., Berking, M., Riper, H., and Ebert, D. D. (2018). Barriers and facilitators for the implementation of blended psychotherapy for depression: A qualitative pilot study of therapists' perspective. *Internet Interv.* 12, 150–164. doi:10.1016/j.invent.2018.01.002.

Triple-E (2017). Definition blended treatment Heleen Riper. Available at: https://www.triple-ehealth.nl/en/projects-cat/depressie/.

Tromp, E. (2011). Multilingual sentiment analysis on social media. *Master's Thesis. Dep. Math. Comput. Sci. Eindhoven Univ. Technol.*, 114. Available at: http://www.win.tue.nl/~mpechen/projects/pdfs/Tromp2011.pdf.

TyranoBuilder (2018). Available at: http://tyranobuilder.com/.

van Ballegooijen, W., Cuijpers, P., Van Straten, A., Karyotaki, E., Andersson, G., Smit, J. H., et al. (2014). Adherence to internet-based and face-to-face cognitive behavioural therapy for depression: A meta-analysis. *PLoS One* 9, e100674. doi:10.1371/journal.pone.0100674.

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., and Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): protocol of a randomised controlled trial. *BMC Psychiatry* 16, 359. doi:10.1186/s12888-016-1065-5.

Van Breda, W. R. J. (2020). Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment. Available at: https://research.vu.nl/en/publications/predictive-modeling-in-e-mental-health-exploring-applicability-in.

Van Breda, W., Treur, J., and van Wissen, A. (2012). "Analysis and Support of Lifestyle via Emotions Using Social Media," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Lecture Notes in Computer Science., eds. K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang, and C. Guéret (Berlin, Heidelberg: Springer Berlin Heidelberg), 275–291. doi:10.1007/978-3-642-35386-4_21.

van der Vaart, R., Witting, M., Riper, H., Kooistra, L., Bohlmeijer, E. T., and van Gemert-Pijnen, L. J. E. W. C. (2014). Blending online therapy into regular face-to-face therapy for depression: Content, ratio and preconditions according to patients and therapists using a Delphi study. *BMC Psychiatry* 14, 1–10. doi:10.1186/s12888-014-0355-z.

Van Vliet, I. M., and De Beurs, E. (2007). Het Mini Internationaal Neuropsychiatrisch Interview (MINI): Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV-en ICD-10-stoornissen. *Tijdschr. Psychiatr.* 49, 393–397.

van Vugt, H. C., Hoorn, J. F., and Konijn, E. A. (2009). Interactive engagement with embodied agents: an empirically validated framework. *Comput. Animat. Virtual Worlds* 20, 195–204. doi:10.1002/cav.312.

van Vugt, H. C., Konijn, E. A., Hoorn, J. F., Keur, I., and Eliëns, A. (2007). Realism is not all! User engagement with task-related interface characters. *Interact. Comput.* 19, 267–280. doi:10.1016/j.intcom.2006.08.005.

Vis, C., Kleiboer, A., Prior, R., Bønes, E., Cavallo, M., Clark, S. A., et al. (2015). Implementing and up-scaling evidence-based eMental health in Europe: The study protocol for the MasterMind project. *Internet Interv.* 2, 399–409. doi:10.1016/j.invent.2015.10.002.

Wainer, J., Dautenhahn, K., Robins, B., and Amirabdollahian, F. (2014a). A Pilot Study with a Novel Setup for Collaborative Play of the Humanoid Robot KASPAR with

Children with Autism. *Int. J. Soc. Robot.* 6, 45–65. doi:10.1007/s12369-013-0195-x.

Wainer, J., Robins, B., Amirabdollahian, F., and Dautenhahn, K. (2014b). Using the Humanoid Robot KASPAR to Autonomously Play Triadic Games and Facilitate Collaborative Play Among Children With Autism. *Ieee Trans. Auton. Ment. Dev.* 6, 183–199. doi:10.1109/tamd.2014.2303116.

Wallace, S., Parsons, S., Westbury, A., White, K. K., White, K. K., and Bailey, A. (2010). Sense of presence and atypical social judgments in immersive virtual environments: Responses of adolescents with Autism Spectrum Disorders. *Autism* 14, 199–213. doi:10.1177/1362361310363283.

Walsh, C. G., Ribeiro, J. D., and Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J. Child Psychol. Psychiatry* 59, 1261–1270. doi:10.1111/jcpp.12916.

Waltz, J., Addis, M. E., Koerner, K., and Jacobson, N. S. (1993). Testing the Integrity of a Psychotherapy Protocol: Assessment of Adherence and Competence. *J. Consult. Clin. Psychol.* 61, 620–630. doi:10.1037/0022-006X.61.4.620.

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry* 14, 270–277. doi:10.1002/wps.20238.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. (2013). "A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network," in, 201–213. doi:10.1007/978-3-642-40319-4_18.

Warmerdam, L., Riper, H., Klein, M., van den Ven, P., Rocha, A., Ricardo Henriques, M., et al. (2012). Innovative ICT solutions to improve treatment outcomes for depression: the ICT4Depression project. *Stud. Health Technol. Inform.* 181, 339–43. doi:10.3233/978-1-61499-121-2-339.

Warren, Z., Zheng, Z., Das, S., Young, E. M., Swanson, A., Weitlauf, A., et al. (2015). Brief Report: Development of a Robotic Intervention Platform for Young Children with ASD. *J. Autism Dev. Disord.* 45, 3870–6. doi:10.1007/s10803-014-2334-0.

Watson, J. C., and McMullen, E. J. (2005). An Examination of Therapist and Client Behavior in High- and Low-Alliance Sessions In Cognitive-Behavioral Therapy and Process Experiential Therapy. *Psychother. Theory, Res. Pract. Train.* 42, 297–310. doi:10.1037/0033-3204.42.3.297.

Watzke, B., Rueddel, H., Koch, U., Rudolph, M., and Schulz, H. (2008). Comparison of therapeutic action, style and content in cognitive-behavioural and psychodynamic group therapy under clinically representative conditions. *Clin. Psychol. Psychother.* 15, 404–417. doi:10.1002/cpp.595.

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

Wentzel, J., van der Vaart, R., Bohlmeijer, E. T., and van Gemert-Pijnen, J. E. W. C. (2016). Mixing Online and Face-to-Face Therapy: How to Benefit From Blended Care in Mental Health Care. *JMIR Ment. Heal.* 3, e9. doi:10.2196/mental.4534.

Wenze, S. J., and Miller, I. W. (2010). Use of ecological momentary assessment in mood disorders research. *Clin. Psychol. Rev.* 30, 794–804. doi:10.1016/j.cpr.2010.06.007.

Wilson, T., Wiebe, J., and Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. *Acl* 7, 12–21. doi:10.3115/1220575.1220619.

Wind, T. R., Rijkeboer, M., Andersson, G., and Riper, H. (2020). The COVID-19 pandemic: The 'black swan' for mental health care and a turning point for e-health. *Internet Interv.*, 100317. doi:10.1016/j.invent.2020.100317.

Wittkampf, K. A., Naeije, L., Schene, A. H., Huyser, J., and van Weert, H. C. (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen. Hosp. Psychiatry* 29, 388–395. doi:https://doi.org/10.1016/j.genhosppsych.2007.06.004.

World Health Organization (2018). Depression fact-sheet. *Depress. fact-sheet*. Available at: https://www.who.int/news-room/fact-sheets/detail/depression [Accessed October 24, 2019].

Yasavur, U., Lisetti, C., and Rishe, N. (2014). Let's talk! speaking virtual counselor offers you a brief intervention. *J. Multimodal User Interfaces* 8, 381–398. doi:10.1007/s12193-014-0169-9.

Zaichkowsky, J. L. (1994). The Personal Involvement Inventory: Reduction, Revision, and Application to Advertising. *J. Advert.* 23, 59–70. doi:10.1080/00913367.1943.10673459.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cybern.* 1, 43–52. doi:10.1007/s13042-010-0001-0.

Zickgraf, H. F., Chambless, D. L., McCarthy, K. S., Gallop, R., Sharpless, B. A., Milrod, B. L., et al. (2016). Interpersonal Factors Are Associated with Lower Therapist Adherence in Cognitive-Behavioural Therapy for Panic Disorder. *Clin. Psychol. Psychother.* 23, 272–284. doi:10.1002/cpp.1955.

Appendices

Appendix A (Chapter 2)

A1 Search strings

PubMed (http://www.ncbi.nlm.nih.gov/pubmed/)

Search String: *("conversational agent" OR "virtual coach" OR "virtual agent" OR "embodied agent" OR avatar OR "relational agent" OR "interactive agent" OR "virtual character" OR "animated character" OR "virtual human" OR humanoid) AND ("mood disorder" OR "affective disorder" OR bipolar OR depression OR depressed OR depressive OR depression* OR depressive* OR melancholia OR anxiety OR agoraphobia OR "obsessive-compulsive disorder" OR "panic disorder" OR "phobic disorder" OR phobia OR "post-traumatic stress disorder" OR "posttraumatic stress" OR ptsd OR anxiety* OR anxious* OR phobi* OR panic* OR schizophrenia OR "psychotic disorder" OR psychoses OR psychosis OR anorexia OR binge-eating OR bulimia OR autism OR autistic OR "substance-related disorder" OR "drug dependence" OR "substance dependence" OR addiction OR "drug abuse" OR "substance abuse" OR alcoholism OR alcohol OR smoking OR tobacco OR cigarette OR nicotine OR cannabis OR marihuana OR marijuana OR "psychotherapy"[Mesh] OR "depressive disorder"[Mesh] OR "substance-related disorders"[Mesh] OR "anxiety disorders"[Mesh] OR "schizophrenia and disorders with psychotic features"[Mesh] OR "autistic disorder"[Mesh])*

WebOfScience (http://webofscience.com/)

Additional Settings: Searched on topic

Search String: *("conversational agent" OR "virtual agent" OR "embodied agent" OR "communicative agent" OR avatar OR "interface agent" OR "relational agent" OR "interactive agent" OR "believable agent" OR "game character" OR "non-playable character" OR "conversational character" OR "believable character" OR "virtual character" OR "synthetic character" OR "animated character" OR "virtual human" OR "virtual coach" OR "virtual therapist" OR "virtual assistant" OR humanoid OR ePartner) AND ("mood disorder" OR "affective disorder" OR bipolar OR depressive OR depression OR depressed OR melancholia OR anxiety OR agoraphobia OR "obsessive-compulsive disorder" OR "panic disorder" OR "phobic disorder" OR phobia OR "post-traumatic stress disorder" OR "posttraumatic stress" OR ptsd OR schizophrenia OR schizophrenic OR psychotic OR psychosis OR psychoses OR "eating disorder" OR anorexia OR bulimia OR binge-eating OR autism OR autistic OR "substance-related disorder" OR "drug dependence" OR "substance dependence" OR addiction OR "drug abuse" OR "substance abuse" OR alcoholism OR alcohol OR smoking OR tobacco OR nicotine OR cigarette OR cannabis OR marihuana OR marijuana)*

ScienceDirect (http://www.sciencedirect.com/)

Search String: *title-abstr-key(("conversational agent" OR "virtual agent" OR "embodied agent" OR "communicative agent" OR avatar OR "interface agent" OR "relational agent" OR "interactive agent" OR "believable agent" OR "game character" OR "non-playable character" OR "conversational character" OR "believable character" OR "virtual character" OR "synthetic character" OR "animated character" OR "virtual human" OR "virtual coach" OR "virtual therapist" OR "virtual assistant" OR humanoid OR ePartner) AND ("mood disorder" OR "affective disorder" OR bipolar OR depressive OR depression OR depressed OR melancholia OR anxiety OR agoraphobia OR "obsessive-compulsive disorder" OR "panic disorder" OR "phobic disorder" OR phobia OR "post-traumatic stress disorder" OR "posttraumatic stress" OR ptsd OR schizophrenia OR schizophrenic OR psychotic OR psychosis OR psychoses OR "eating disorder" OR anorexia OR bulimia OR binge-eating OR autism OR autistic OR "substance-related disorder" OR "drug dependence" OR "substance dependence" OR addiction OR "drug abuse" OR "substance abuse" OR alcoholism OR alcohol OR smoking OR tobacco OR nicotine OR cigarette OR cannabis OR marihuana OR marijuana))*

SpringerLink (http://springerlink.com)

Additional Settings: results filtered by HCI and AI

Search String: see WebOfScience

ACM Digital Library (http://dl.acm.org/)

Search String: *((Abstract:"conversational agent" or Abstract:"virtual agent" or Abstract:"embodied agent" or Abstract:"communicative agent" or Abstract:avatar or Abstract:"interface agent" or Abstract:"relational agent" or Abstract:"interactive agent" or Abstract:"believable agent" or Abstract:"game character" or Abstract:"non-playable character") and (Abstract:"mood disorder" or Abstract:"affective disorder" or Abstract:bipolar or Abstract:depressive or Abstract:depression or Abstract:depressed or Abstract:melancholia or Abstract:anxiety or Abstract:agoraphobia or Abstract:"obsessive-compulsive disorder" or Abstract:"panic disorder" or Abstract:"phobic disorder" or Abstract:phobia or Abstract:"post-traumatic stress disorder" or Abstract:"posttraumatic stress" or Abstract:ptsd or Abstract:schizophrenia or Abstract:schizophrenic or Abstract:psychotic or Abstract:psychosis or Abstract:psychoses or Abstract:"eating disorder" or Abstract:anorexia or Abstract:bulimia or Abstract:binge-eating or Abstract:autism or Abstract:autistic or Abstract:"substance-related disorder" or Abstract:"drug dependence" or Abstract:"substance dependence" or Abstract:addiction or Abstract:"drug abuse" or Abstract:"substance abuse" or Abstract:alcoholism or Abstract:alcohol or*

Abstract:smoking or Abstract:tobacco or Abstract:nicotine or Abstract:cigarette or Abstract:cannabis or Abstract:marihuana or Abstract:marijuana))

((Title:"conversational character" or Title:"believable character" or Title:"virtual character" or Title:"synthetic character" or Title:"animated character" or Title:"virtual human" or Title:"virtual coach" or Title:"virtual therapist" or Title:"virtual assistant" or Title:humanoid or Title:ePartner) and (Title:"mood disorder" or Title:"affective disorder" or Title:bipolar or Title:depressive or Title:depression or Title:depressed or Title:melancholia or Title:anxiety or Title:agoraphobia or Title:"obsessive-compulsive disorder" or Title:"panic disorder" or Title:"phobic disorder" or Title:phobia or Title:"post-traumatic stress disorder" or Title:"posttraumatic stress" or Title:ptsd or Title:schizophrenia or Title:schizophrenic or Title:psychotic or Title:psychosis or Title:psychoses or Title:"eating disorder" or Title:anorexia or Title:bulimia or Title:binge-eating or Title:autism or Title:autistic or Title:"substance-related disorder" or Title:"drug dependence" or Title:"substance dependence" or Title:addiction or Title:"drug abuse" or Title:"substance abuse" or Title:alcoholism or Title:alcohol or Title:smoking or Title:tobacco or Title:nicotine or Title:cigarette or Title:cannabis or Title:marihuana or Title:marijuana))

A2 Concept definitions

Concept	Definition
Meta-Information	
Authors	The authors of the study.
Year of Publication	The year in which the study was published.
Publication Medium	The medium in which the study was published, e.g., conference proceedings, journal, or book series.
Study Characteristics	
Primary Institute	The research institute of the primary author.
Country	The country in which the primary institute is located.
Project Name	Name of the project or research group.
Targeted Disorder	The disorder that was targeted in the study.
Intended Intervention	
Social skills training	Improving the user's social skills
CBT	Changing cognitions or behaviors based on CBT-principles
Counseling	Getting users to communicate about their problems in a therapeutic dialogue
Educational aid	Contributing to the user's knowledge, not related to the disorder
Self-management	Stimulating the user's application of behavior change tactics to produce desired changes
Target Skill or Behavior	The specific skill or behavior that the intervention aimed to improve or change.
Number of Participants	The number of participants in the study.
Age Category	
preschoolers	age 0-4
children	age 5-12
adolescents	age 13-17
young adults	age 18-24
adults	age 25-64
elderly	age 65 and older

Recruitment Setting	
educational	Participants were recruited through schools or universities.
clinical	Participants were recruited through health-care facilities.
community	Participants were recruited otherwise.
Clinical Population	Whether or not the participants were from the targeted clinical population.
Independent Diagnosis	Whether or not the above was verified through a reported independent diagnosis.
Study Methodology	
Evaluation criterion	
ECA parameters	The study evaluated particular parameter settings of the ECA (e.g., ECA gender, or a mental model).
ECA	The study evaluated whether or not the ECA itself was of additional benefit.
intervention	The study evaluated an entire intervention that included an ECA.
RCT	Were participants randomized in a treatment and control group?
Outcome Type	
user satisfaction	Whether or not users respond positively to the intervention.
usability	Whether or not users have trouble using the intervention.
usage	How often and how the intervention is used.
behavioral	Whether or not the study assesses user behavior.
knowledge	Whether or not users acquired targeted knowledge by using the intervention.
self-report	When users or their care-givers report on their own experience/behavior.
Study Aim	What the study aimed to find out.
Main Conclusion	The main conclusion of the study.
Development Phase	
development	The intervention is still subject to changes, and measures are related to usability, satisfy action, and feasibility. Measures do not yet include thorough evaluation based on relevant clinical outcomes. These types of studies describe interesting new developments that require further evaluation.

piloting	The intervention is near completion, and relevant patient outcomes are taken into account in the evaluation. Usability, satisfaction, and feasibility outcomes can go hand-in-hand with patient outcomes. Evidence is not yet significant enough to give clinicians enough confidence to apply it in practice.
evaluation	The evaluation revolves primarily around the intervention's effect on clinical outcomes. Sample sizes are typically larger, and methodology is more rigorous. These interventions have been evaluated to the extent that clinicians could consider their practical application.
implementation	The intervention has already gone through the evaluation phase, and has been used in practice for some time.
ECA Characteristics	
Platform	The platform in which the ECA was embodied.
serious game	A game with a primary purpose other than entertainment
stand-alone software	Platforms that do not belong to any of the other categories
robotics	Platforms with a physical embodiment rather than a virtual one
virtual reality	Truly immersive (more so than just a computer screen) applications
web-based	Applications that run in a web-browser
Platform Specification	A further specification of the platform if applicable.
Personification	Who or what the ECA embodiment personifies.
Personification Specification	A further specification of the personification if applicable.
Social Role	
Social interaction partner	To engage in an interaction with the user to improve specific social skills
Tutor	To teach something to the user
Coach	To motivate and engage the user
Health-care provider	To simulate the behavior of a health-care provider
Purpose	What it is that the ECA does in the application.
Dialog	Whether or not users could enter into a dialog with the ECA.
Dialog Specification	How users could enter into the dialog.

Human-ECA Interaction	How users could communicate with the ECA.
Human Communication Modalities Used	Whether this involved human communication modalities.
Communication Modalities	If so, which communication modalities.
ECA-Human Interaction	How the ECA communicated with the user.
Human Communication Modalities Used	Whether this involved human communication modalities.
Communication Modalities	If so, which communication modalities
User Model	Whether the ECA kept track of information on the user to personalize the interactions.
static	the model was based on information entered before the interactions (e.g., participant age)
dynamic	the model was updated throughout the interactions to keep track of a user's changing states (e.g., emotional state)

A3 Overview of the intervention aims and ECA characteristics

Disorder / Author	Intervention	Intervention targets	Platform	ECA Embodiment	ECA Social Role
Autism					
Agarwal (2013)	social skills training	turn-taking, joint attention	game	fantasy, smiley	social interaction partner
Alcorn (2011)	social skills training	joint attention	game	human	social interaction partner
Amirabdollahian (2011)	social skills training	tactile interaction	robotics	humanoid	social interaction partner
Bamasak (2013)	educational aid	cognitive, social, and self-care skills	game	human	tutor
Bekele (2013) Bekele (2014)	social skills training	joint attention	robotics	humanoid	social interaction partner
Bernardini (2012) Bernardini (2014)	social skills training	joint attention	game	human	social interaction partner
Boccanfuso (2010)	social skills training	turn-taking, imitation	robotics	toy-like	social interaction partner
Chen (2010)	social skills training	joint attention	game	human	social interaction partner
Cole (2003)	educational aid	vocabulary	stand-alone	human	tutor
Costa (2015)	social skills training	tactile interaction, body consciousness	robotics	humanoid	social interaction partner
Dickerson (2013)	social skills training	tactile interaction	robotics	humanoid	social interaction partner

Fujimoto (2010) Fujimoto (2011)	social skills training	imitation	robotics	humanoid	social interaction partner
Hopkins (2011)	social skills training	joint attention, facial & emotion rec	game	photograph	social interaction partner, coach
Jordan (2013)	social skills training	communication, turn-taking	robotics	toy-like	social interaction partner
Kim (2010)	social skills training	positive play	robotics	toy-like	social interaction partner
Konstantinidis (2009)	educational aid	special needs	stand-alone	human	tutor
Lahiri (2011)	social skills training	communication	virtual reality	human	social interaction partner
Milne (2009)	social skills training	communication	stand-alone	human	tutor
Palestra (2014)	educational aid	body consciousness	robotics	humanoid	tutor
Ribeiro (2014)	social skills training	communication	game	human	tutor
Robins (2014)	social skills training	tactile interaction, imitation	robotics	humanoid	social interaction partner
Shoukry (2015)	educational aid	learning idioms	game	human	tutor
Smith (2014a)	social skills training	job interview skills	web-based	human	social interaction partner, coach
Tanaka (2015)	social skills training	communication	stand-alone	human	tutor
Wainer (2014a) Wainer (2014b)	social skills training	cooperation	robotics	humanoid	social interaction partner

Warren (2014)	social skills training	imitation	robotics	humanoid	social interaction partner
Depression					
Bickmore (2010b)	self-management	hospital discharge	stand-alone	human	health-care provider
Cheek (2014)	CBT	symptoms of depression	game	fantasy	coach
Kelders (2015)	CBT	symptoms of depression	web-based	photograph	health-care provider
Martínez-Miranda (2014)	CBT	symptoms of depression	web-based	human	coach
Pagliari (2012)	CBT	symptoms of depression	web-based	unknown	coach
Pinto (2013) Pinto (2015)	self-management	health communication	game	human	health-care provider, coach
Pontier (2008)	counseling	clinical assessment	web-based	human	health-care provider
DeVault (2014)[a]	counseling	self-disclosure	stand-alone	human	health-care provider
Swartout (2013)[b]	counseling	screening	web-based	human	coach
Smith (2014b)[c]	social skills training	job interview skills	web-based	human	social interaction partner, coach
Anxiety					
Kang (2010)	counseling	self-disclosure	stand-alone	human	social interaction partner
Kang (2012)	counseling	self-disclosure	stand-alone	human	health-care provider
Rinck (2010)	CBT	exposure to social situations	virtual reality	human	social interaction partner

Schmidt (2013)	CBT	performance anxiety	game	animal	social interaction partner, tutor
PTSD					
Morie (2009)	self-management	symptoms of PTSD	game	fantasy	coach
Tielman (2014)	CBT	memory restructuring	web-based	unknown	coach
Schizophrenia					
Bickmore (2010c)	self-management	medication adherence	stand-alone	human	health-care provider
Ku (2007)	social skills training	conversation skills	virtual reality	human	social interaction partner, tutor
Puskar (2011)	self-management	medication adherence	stand-alone	human	health-care provider
Substance Abuse					
An (2013)	CBT	smoking cessation	web-based	human	coach
Grolleman (2006)	CBT	smoking cessation	web-based	human	coach
Lisetti (2013)	CBT	alcohol dependency	web-based	human	health-care provider
Yasavur (2014)	CBT	alcohol dependency	web-based	human	health-care provider

CBT = Cognitive Behavioral Therapy; ECA = Embodied Conversational Agent; PTSD = Post-Traumatic Stress Disorder

[a] *also targeted anxiety and PTSD*

[b] *also targeted PTSD*

[c] *also targeted schizophrenia*

A4 Overview of study design characteristics by disorder

Disorder / Author	N	Target Group	Clinical Sample	Outcomes	Development Phase
Autism					
Agarwal (2013)	4	children, adolescents	yes	usability	development
Alcorn (2011)	32	children, adolescents	yes	usability, behavioral	development
Amirabdollahian (2011)	-	preschoolers	yes	usage	development
Bamasak (2013)	10	children, adolescents	yes	usability	development
Bekele (2013) Bekele (2014)	12	preschooler	yes	usability, usage, behavioral	piloting
Bernardini (2012) Bernardini (2014)	19	preschoolers, children, adolescents	yes	behavioral, usage	piloting
Boccanfuso (2010)	-	children	no	usage	development
Chen (2010)	-	children	yes	usage	development
Cole (2003)	6	preschoolers, children	yes	knowledge, behavioral	piloting
Costa (2015)	8	children	yes	behavioral, knowledge, usage	piloting
Dickerson (2013)	1	children	yes	behavioral	piloting
Fujimoto (2010) Fujimoto (2011)	4	children	yes	behavioral, usage	development
Hopkins (2011)	49	children, adolescents	yes	behavioral, self-report	evaluation
Jordan (2013)	5	young adults	yes	behavioral, usage	piloting
Kim (2010)	3	preschooler	yes	behavioral	development
Konstantinidis (2009)	13	adults	no	self-report, usability	development

Lahiri (2011)	4	adolescents	yes	behavioral	development
Milne (2009)	7	children, adolescents	unknown	knowledge, usability	development
Palestra (2014)	10	young adults	no	satisfaction, usability	development
Ribeiro (2014)	4	children	yes	behavioral	piloting
Robins (2014)	36	preschoolers, children	no	behavioral, usability	piloting
Shoukry (2015)	12	children, adolescents	yes	usability, knowledge, usage	development
Smith (2014a)	26	adults	yes	usability, usage, satisfaction, behavioral	evaluation
Tanaka (2015)	49	young adults	no	behavioral, usability, satisfaction	piloting
Wainer (2014a) Wainer (2014b)	6	unknown	yes	behavioral	piloting
Warren (2014)	16	preschoolers	yes	behavioral, usage	piloting
Depression					
Bickmore (2010b)	131	young adults, adults	yes	satisfaction, usability, self-report	piloting
Cheek (2014)	16	adolescents	yes	satisfaction	implementatio n
Kelders (2015)	134	adults	yes	behavioral, self-report	evaluation
Martínez-Miranda (2014)	8	young adults, adults	yes	satisfaction	development
Pagliari (2012)	-	unknown	yes	unknown	development
Pinto (2013) Pinto (2015)	28	young adults	yes	self-report, satisfaction	piloting
Pontier (2008)	28	unknown	no	satisfaction	development

DeVault (2014)[a]	351	adults	no	satisfaction, usability, self-report	development
Swartout (2013)[b]	111	adults	no	usability, satisfaction	development
Smith (2014b)[c]	37	adults	yes	usability, usage, satisfaction, behavioral	evaluation
Anxiety					
Kang (2010)	108	adults	no	behavioral	piloting
Kang (2012)	40	adults	no	behavioral	piloting
Rinck (2010)	23	young adults	no	self-report, behavioral	piloting
Schmidt (2013)	15	children	no	usability, self-report, usage	development
PTSD					
Morie (2009)	700	adults	yes	usability	development
Tielman (2014)	10	adults	no	unknown	development
Schizophrenia					
Bickmore (2010c)	15	young adults, adults	yes	satisfaction, usage, self-report	piloting
Ku (2007)	10	adults	yes	satisfaction, usability, behavioral, self-report	piloting
Puskar (2011)	17	young adults, adults	yes	behavioral, satisfaction, usability	piloting
Substance Abuse					
An (2013)	1317	young adults, adults	yes	self-report	evaluation
Grolleman (2006)	35	young adults	yes	unknown	development
Lisetti (2013)	81	young adults	no	usability, satisfaction, self-report	piloting

Yasavur (2014)	89	young adults	no	usability, satisfaction, usage	Piloting

'-' indicates that sample sizes were not reported.

[a] *also targeted anxiety and PTSD*

[b] *also targeted PTSD*

[c] *also targeted schizophrenia*

Appendix B (Chapter 5)

B1 Main and subcategories therapist behaviors

Category therapist behaviors	Definition	Example	%
Emphasizing responsibility	Emphasize the responsibility of the pt[a]	*It is useful for yourself to get insight in this by filling in the diary*	0.1
Affirming	Give attention/recognize/show interest in thoughts, emotions and behaviors of the pt and to consider them valid		22.2
Validating and interpreting	Confirm by interpreting and validating what the pt has written	*That must be very difficult for you I see that you are struggling with it*	10.9
Normalizing	Confirm by stating that the behavior of the pt often occurs/is normal	*It is not easy to .... /It is very common that you feel this way*	1.4
Summarizing online	Confirm by summarizing and repeating what the pt has written	*You write that you..../You are able to focus and enjoy your activities*	7.4
Summarizing f2f	Confirm by summarizing and repeating what the pt has said	*This morning we discussed that you....*	2.5
Clarifying the framework	Clarify/emphasize/remind the pt of the protocol/framework and give practical information	*You gave now finished the first part of the treatment, in the second part we will focus on....*	3.9
Self-disclosure	Use experiences and personal examples from the therapist's life	*This exercise has helped me with my sleeping problems*	0
Informing	Inform or refer to different functionalities on the online platform		27.5
Informing about the next session	Inform or refer to the next session and/or its content	*You can now proceed with session 5: Structure and planning*	10.4
Informing about the diary	Inform or refer to the diary and/or its content	*It is helpful to set a reminder in your phone to fill in the dairy everyday*	2.8
Informing about the monitoring	Inform or refer to the monitoring/ questionnaire and/or its content	*Your score on the QIDS is 7, so your symptoms are mild now*	1.5
Informing about the assignments	Inform or refer to the assignments in the previous session and/or its content	*You kept track of your activities during the weekend, I'm curious about your week-days*	3.7
Informing about the f2f session	Inform or refer to face-to-face session	*See you next week, at 14.00!*	9.1
Confronting	Express a different opinion or disagree with the pt	*I do not think that your activities are useless, but I think the opposite!*	0.4
Urging	Urge to let the pt do something	*It is important that you try this*	3.9

Encouraging	Encourage/motivate past and future behavior of the pt		23.4
Praising past behavior	Praise something the pt has done in the past	*Even though you found it difficult, you did it, very good!*	14.5
Inciting future behavior	Incite something that the pt is planning to do	*Continue to keep up this good work! Good luck!*	8.9
Guiding	Give advice, information or suggestions		11.3
Psychoeducation	Give information on psychological processes	*The fact that it doesn't immediately give more pleasure is something we see often in a depression*	5.5
Giving suggestions	Give suggestions about alternative behavior/provide advice on how it can be addressed	*Try to think of something that's easy to do and cannot easily go wrong*	5.8
Questions	Ask or answer questions		6.9
Asking questions to clarify	Ask questions to better understand the behavior or emotions of the pt	*Is that something you already do?/ … is that correct?*	3.4
Asking 'thought' questions	Ask questions to encourage the pt to think further	*How would you like to feel and what should your life look like?*	3.2
Answering questions	Respond to questions of the pt	*Yes, I will repeat this session for you, so that you can practice more*	0.3

[a] *Pt stands for patient.*

B2 Main and subcategories feedback instructions

Category feedback instructions	Description in instructions	Example	Yes %
Greeting/ending			**95.9**
Correct greeting	Use a greeting	*Dear [name pt[a]]*	95.4
Correct ending	Use a correct ending	*Good luck with the next session! Greetings, [name therapist]*	96.3
Communication skills			**69.4**
Begin with compliment	Begin the message with a compliment/positive approach	*From your answers I can see it was difficult, but you did very good!*	79.0
Summarize	Summarize the assignments of the pt, use his/her own words	*You clearly described your depression, the loss of interest/energy, negative thoughts and worrying.*	73.1
Reading homework	Show that you have read the homework by using their examples	*I see that you.../You say that you..../If I understand you correctly...*	88.6
Hypotheses	Formulate sentences as hypotheses	*That sounds like..... , is that correct?.. If I understand you correctly.....*	10.5
Giving no solutions	Give no solutions, let patients think of their own solutions. You can give them a starter in the right direction	*When was the last time you felt that way? What did you think and what did you do differently then?*	95.6
Structure			**87.7**
2 subjects	Give feedback on max two subjects in order to keep the structure clear		95.4
Within 3 working days	Give feedback within three working days. Make an agreement when the pt should finish the session and when you as a therapist can write the feedback		79.9
Referring			**34.0**
Referring diary	Reflect on the diary or when the pt did not succeed yet, address (again) the utility of the diary	*Good that you used the diary. I can see that your mood is different in the mornings than in the evenings, it that something you also experience?*	26.9
Referring monitoring	Refer to the monitoring questionnaire or when the pt did not succeed yet, ask to fill this in next time	*Your score on the QIDS is 7, so right now your symptoms are mild.*	11.4

Referring next session	End the message with a reference to the content of the next session	*I will open the next session for you. There you can further practice with the schemas*	41.6
Referring next f2f session	Confirm the next f2f appointment	*We will see each other again on May 19.*	56.2
Readability			**68.0**
Short sentences	Write short, clear sentences.		61.6
Short paragraphs	Write short, clear paragraphs		74.4
Writing style			**93.6**
Limit abbreviations	Limit your abbreviations		97.3
Limit misspellings	Be aware of misspellings		78.5
Limit emphasis	Limit text in caps, exclamation marks, underline text or make text bold		98.6
Correct emoticons	Use emoticons only when you know the meaning is clear for the pt and use it as a complement and not as a substation	*And do not forget the fun activities ☺*	100

[a] *Pt stands for patient.*

B3 Case descriptions of 3 patients

B3.1 Case description of patient K

B3.1.1 Background information

K. is a 40 year old father of one child. He works as a chef. He suffers from fatigue, feelings of exhaustion, anhedonia, emotional instability and excessive guilt. His symptoms seem to be caused by an accumulation of stressors over the last few years (high workload, illness of his mother, job change of partner and worrying about his daughter), in combination with a habit of making high demands on himself and a need to always be strong. This is the third time he is experiencing these symptoms. There seems to be an overload of stress for some time now, which has now led to the development of a depressive disorder.

B3.1.2 Treatment course

K. completed 11 online sessions (repeated online session 3) and had 5 face-to-face sessions with his therapist. At the beginning of treatment K. scores 17 on the QIDS, which is considered severe. The treatment was successful, after 15 weeks of treatment the depression was in remission (QIDS score 7, severity mild). Most used therapist behaviors were Affirming (relative frequency 24.7), Informing (relative frequency 19.7) and Encouraging (relative frequency 18,3).

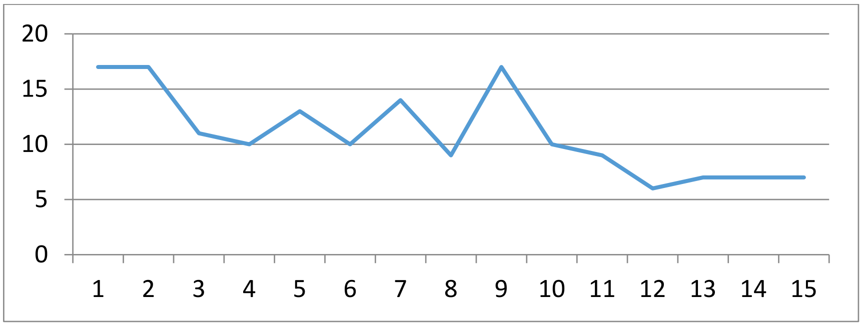

Figure 1. QIDS scores of patient K.

Table 1. Quotes and therapist behavior codes of online sessions of patient K.

Online sessions	Quotes from feedback messages	Codes
Online session 2 'Clarifying your symptoms'	Feeling guilty, is that something you always do (to a greater or lesser extent)? Or do you suffer from it now, during your depression? (In any case, it is also one of the symptoms of depression, but it can also be something that you do anyway, and then it is probably a long-standing thought pattern.	asking questions to clarify, psychoeducation
Online session 3 'Motivation and setting goals'	We have just discussed that you will repeat this session again. You will briefly describe the day, so that you can get concrete goals for the coming period. What do you want to achieve with the therapy? Make that concrete and think about one (or more) steps that you can take. We just discussed that it might be wise, for example, that you give your child more responsibility. You do not have to do everything for her. Children are allowed to do tasks, that is also part of parenting.	summarizing f2f, asking 'thought' questions, giving suggestions, normalizing
Online session 10 'Looking at the future'	You have achieved good results with the therapy. You can enjoy time with your child and friends again and you realize how this happened by structurally taking too much hay on your fork. You now know how it feels to be rested. You have gained many insights by critically examining your own thoughts. Keep challenging your thoughts, so that your new thoughts / rules of life will become increasingly credible. Your score on QIDS = 7, so still mild complaints (0-5 = no depressive symptoms). This means that you experience a lot of improvement, the symptoms you have are still light and I expect if you persevere, that you can maintain the improvement	summarizing online, inciting future behavior, informing about the monitoring, validating/interpreting

B3.2 Case description of patient A

B3.2.1 Background information

A. is 25 years old and experiences problems with her work. Her symptoms have been there for two to three years now. She has a negative self-image and shows a significant amount of avoidance behavior. She sometimes thinks of suicide, but says she doesn't want to die. Her family lives in Turkey.

B3.2.2 Treatment course

A. completed 8 online sessions (repeated online session 5) and had 5 face-to-face sessions with her therapist. At the beginning of treatment she scores 18 on the QIDS, which is considered severe. After 17 weeks of treatment A. is still experiencing severe symptoms of depression and she scores 21 on the QIDS. Despite this, she indicates at the end of treatment that she has gained a lot of insight from the treatment. Apart from the online sessions, she also received guidance from a self-help book about a negative self-image. After treatment she wants to continue to use the online thought-schemes. Most used therapist behaviors were Informing (relative frequency 31.6), Affirming (relative frequency 21.1) and Encouraging (relative frequency 19.3).

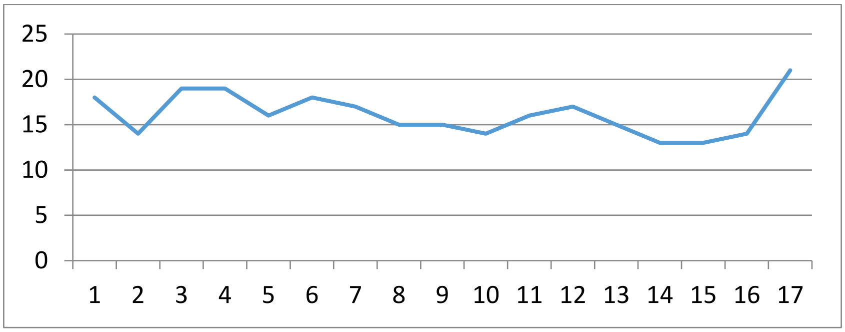

Figure 2. QIDS scores of patient A.

Table 2. Quotes and therapist behavior codes of online sessions of patient A.

Online sessions	Quotes from feedback messages	Codes
Online session 2 'Clarifying your symptoms'	Very nice that you managed to go through the first session and that you succeeded in completing the exercise. Good to hear that you have experienced it as helpful to put everything in order. Good that you filled in the questionnaires! We will discuss them later. I see that you also kept the diary for one day. It is important to do this consistently so that we get an idea of what the link is between your activities and the symptoms you have.	praising past behavior, informing about the monitoring, diary, urging
Online session 5 'Pleasure and fulfilment'	I see that thinking of enjoyable activities goes well, but that you find it difficult to implement them. We have already discussed this in the face to face sessions several times. It's good that you've tried it. You say that you want to do the session again. Shall we agree that you will do it one more time and then actually carry out the intended activity? Try to think of something	validating/interpreting, summarizing f2f, praising past behavior, giving suggestions, urging, asking 'thought' questions

	that is easy to carry out and that cannot go wrong. Whatever happens, allow yourself this activity. Maybe you can also think of a way to reward yourself when you succeed?	
Online session 6 'Structure and planning'	You indicated earlier that you benefited from previous structures and planning for work. So it is worth to try to incorporate this in your private activities as much as possible as well. Try to make a planning for each day by using the steps: Step 1: Plan your every-day activities, Step 2: Plan your must-activities, Step 3: Plan your enjoyable activities. Good luck with the next module about the thought-scheme!	summarizing f2f, informing about the assignments, giving suggestions, inciting future behavior

B3.3 Case description of patient M

B3.3.1 Background information

M. is 25 years old. His depressive symptoms are associated with emotional neglect by his parents. M. is very driven and ambitious, in combination with a vulnerable and low self-esteem. He supports his parents wherever he can, especially financially, but doesn't get recognition for this. M. also suffers from concentration and reading problems following an accident.

B3.3.2 Treatment course

M. completed 4 online sessions and had 8 face-to-face sessions with his therapist. At the beginning of treatment he scores 19 on the QIDS, which is considered severe. At the last known QIDS he scores 11. M. indicates that the therapy has helped him to get more insight into the dynamics within de family. He feels less responsible for his parents. His mood has improved. His self-image and knowledge about his needs remain diffuse. Most used therapist behaviors were Encouraging (relative frequency 28.6), Affirming (relative frequency 21.4) and Questions (relative frequency 21.4).

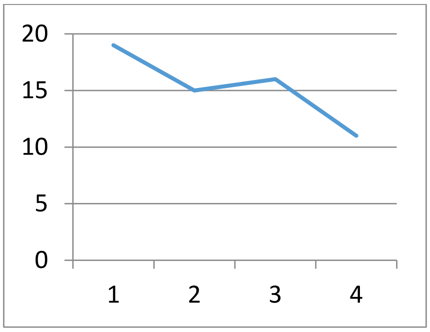

Figure 3. QIDS scores of patient M.

Table 3. Quotes and therapist behavior codes of online sessions of patient M.

Online sessions	Quotes from feedback messages	Codes
Online session 2 'Clarifying your symptoms'	I see you've been working well on the assignment. What was it like to write about your depressive symptoms? It is indeed difficult if you are not in a good mood to do things. And that if you manage to do things, that this will have a positive influence on your mood.	praising past behavior, asking 'thought' questions, validating/interpreting
Online session 3 'Motivation and setting goals'	Thank you for completing the assignments. Well done. You write that is the first step is very difficult, since you want to be less busy, but that means more free time and peace and that turns into thinking and worrying. But is this then a good reason not to take a rest? What could you do about the rumination?	praising past behavior, summarizing online, confronting, asking 'thought' questions
Online session 4 'How active are you?'	Good that you have found a way to deal with this exercise. Has it helped to get better insight between the relationship of activity and mood? You can find more helpful questions in the assignment. Good luck.	praising past behavior, asking questions to clarify, informing about the assignments, inciting future behavior

Appendix C (Chapter 6)

C1 Content of the sessions in the blended CBT protocol

Diagnosis	Session (modality)	Content and exercises
Panic disorder	*1 (f2f)*	Introduction, psychoeducation
	2 (online)	Treatment rationale, treatment motivation and expectation, worry diary
	3 (f2f)	Explanation of exposure, instructions regarding interoceptive exposure, panic diary
	4 (online)	Interoceptive exposure exercises, exposure diary
	5 (f2f)	Explanation of interoceptive exposure and exposure in vivo, exposure exercise, exposure diary
	6 (online)	Identifying automatic thoughts, challenging unhelpful thoughts, exposure diary
	7 (f2f)	Explanation of behavioural experiments, treatment evaluation, exposure diary
	8 (online)	Identifying cognitive distortions, behavioural experiment, exposure diary
	9 (f2f)	Behavioural experiment, exposure diary
	10 (online)	Exposure exercises, behavioural experiments
	11 (f2f)	Exposure exercises, behavioural experiments
	12 (online)	Exposure exercises, behavioural experiments
	13 (f2f)	Exposure exercises, behavioural experiments
	14 (online)	Explanation of relapse prevention, relapse prevention plan
	15 (f2f)	Relapse prevention plan, recapitulation, treatment evaluation
Social anxiety disorder	*1 (f2f)*	Introduction, psychoeducation
	2 (online)	Treatment rationale, treatment motivation and expectation, worry diary
	3 (f2f)	Explanation of selective attention, explanation and instructions regarding exposure, social anxiety diary
	4 (online)	Identifying automatic thoughts, challenging unhelpful thoughts, social anxiety diary
	5 (f2f)	Identifying automatic thoughts, challenging unhelpful thoughts, social anxiety diary
	6 (online)	Identifying cognitive distortions, social anxiety diary

	7 (f2f)	Explanation of behavioural experiments, treatment evaluation
	8 (online)	Behavioural experiments, social anxiety diary
	9 (f2f)	Behavioural experiments, social anxiety diary
	10 (online)	Behavioural experiments, social anxiety diary
	11 (f2f)	Behavioural experiments, social anxiety diary
	12 (online)	Behavioural experiments, social anxiety diary
	13 (f2f)	Behavioural experiments, social anxiety diary
	14 (online)	Explanation of relapse prevention, relapse prevention plan
	15 (f2f)	Relapse prevention plan, recapitulation, treatment evaluation
Generalised anxiety disorder	*1 (f2f)*	Introduction, psychoeducation
	2 (online)	Treatment rationale, treatment motivation and expectation, worry diary
	3 (f2f)	Explanation and instructions regarding exposure, challenging unhelpful thoughts, explanation of metacognitions
	4 (online)	Exploring metacognitions, worry exposure exercises
	5 (f2f)	Exploring uncontrollability of worrying
	6 (online)	Worry experiment regarding uncontrollability
	7 (f2f)	Exploring the danger of worrying, treatment evaluation
	8 (online)	Worry experiment regarding danger of worrying
	9 (f2f)	Exploring positive beliefs about worrying
	10 (online)	Worry experiment regarding positive beliefs
	11 (f2f)	Explanation of selective attention
	12 (online)	Learning to shift attention
	13 (f2f)	Learning to shift attention
	14 (online)	Explanation of relapse prevention, relapse prevention plan
	15 (f2f)	Relapse prevention plan, recapitulation, treatment evaluation

C2 Checklist for fidelity to blended treatment protocol

C2.1 Face-to-face sessions

For face-to-face protocol components, a score of 2 indicates full adherence, 1 partial adherence, and 0 non-adherence.

Protocol component	Checklist	Examples from transcripts
Psychoeducation	**Adherence:** therapist gives a clear explanation of the blended format and logs into the online platform with the patient	Session 1 (panic disorder, PD) **Therapist:** *This treatment will last 15 weeks and face-to-face sessions will alternate with online sessions.* … **Therapist:** *Let's take a look at the online platform.* **Patient:** *Okay.* **Therapist:** *It's really easy, so you'll be fine. Here you see the sessions and your progress, so you know where to start when you log in again. Now we'll have a look at the introduction session together.* **Patient:** *Yes, I can see that in Tasks.* **Therapist:** *That's right, you can see the introduction sessions there. And here you see what the content of the session is.* **Patient:** *I see.*
	Partial adherence: therapist does not give a clear explanation of the blended format *or* does not log into the online platform with patient	
	Non-adherence: therapist does not give a clear explanation of the blended format *and* does not log into the online platform with patient	
Discussing previous online session	**Adherence:** content of homework and exercises in the previous online session is discussed	Session 5 (PD) **Therapist:** *I gave you some feedback on the online session. You had made a list of exposure activities and described your catastrophic thought. You described what happens when you get into your car very clearly. You think: "Oh no, I will stop at the next petrol station."*

		Patient: *That's right.* **Therapist:** *I think those are reactions to your catastrophic thought. It's not that thought that makes you feel anxious and that makes you want to stop driving the car. Because … what might happen if you keep driving? What's the most catastrophic thing that could happen?*
	Partial adherence: reference is made to the online sessions, but content of homework and exercises is not discussed	Session 3 (PD) **Therapist:** *First I want to briefly discuss the online session. How did that go?* **Patient:** *I recognised some things, some of these exercises I've done before. And um … I found it difficult to describe my own situation.* **Therapist:** *Yes.* **Patient:** *That's still difficult for me. The rest of it was clear. When you start working on it, you start thinking about your situation. That's good, I think….* **Therapist:** *I read it and gave you feedback. How did you feel about that?* **Patient:** *I found it supportive.* **Therapist:** *Good.*
	Non-adherence: homework and exercises in the previous online sessions are not discussed	
Preparing upcoming online session	**Adherence:** homework for the upcoming online session is discussed and an appointment for providing feedback is made	Session 3 (PD) **Therapist:** *This week you'll keep your panic diary; you can do that online. That means you describe every panic attack in your diary. For example, if you have trouble breathing again and that makes you feel anxious.* … **Therapist:** *I'll send you feedback on the online session next Monday, in the morning. Is that okay?* **Patient:** *Yes.*
	Partial adherence: homework for the upcoming online session is not discussed *or* no appointment for providing feedback on the online sessions is made	Session 5 (PD) **Therapist:** *We'll see each other again on Wednesday in two weeks' time.* **Patient:** *Yes, that means I can do the exposure activity at least five times.*

	Therapist: *That's good, you can practise a lot.* → An appointment for the next f2f session is made, but not one for providing feedback on the online session.
Non-adherence: homework for the upcoming online sessions is not discussed *and* no appointment for providing feedback on the online sessions is made	

C2.2 Online Sessions

For online protocol components, a score of 1 indicates presence of the component, and 0 indicates non-presence.

Protocol component	Checklist	Examples from feedback messages
Generic therapeutic feedback	Encouraging and motivating	• *I see that you worked very hard!* • *Good job!* • *Well done!*
	Normalising and empathising	• *I can imagine that made you sad and anxious.* • *I understand that it feels like a disappointment.* • *I can imagine that is not easy.* • *Simply put, people feel anxiety in their bodies sometimes, and especially people who are recovering from an anxiety disorder will regularly feel restless and anxious. That can't be prevented.* • *Be aware that you don't have to understand and be able to perform things perfectly at once. I've never met anyone who could (myself included), so it's completely fine if you want to reassure yourself or need more time.* • *The goal was not (yet) to sit on your bike without fear (or fearful sensations). Because indeed, as you write, it takes time to get used to it, also for your body and your mind.*
	Confirming by summarising	• *The most important part of the treatment is indeed to investigate your thoughts.*
	Guiding treatment progress	• *You can log your panic attacks in the panic diary.*

		• *This week's homework is that you will perform three activities from your exposure list.* • *Once you've completed an exercise, it's good to log that in the exposure diary.*
CBT-specific feedback	Explaining CBT theory (including exposure)	• *Often you start by feeling something in your body: your heart starts beating faster, you start hyperventilating, and then you start feeling light-headed. Because you feel light-headed, you start thinking something bad could happen, for example fainting or having a heart attack. Then the reasoning becomes: If my heart beats faster, I will start hyperventilating, I'll become light-headed, and then I'll faint or have a heart attack.* • *The fear of fear became stronger, which made you end up in the panic cycle: anxiety → catastrophic thoughts about anxiety → anxiety increases → catastrophic thoughts become stronger → anxiety increases etc.* • *With an alternative thought we mean a thought that is more realistic than a catastrophic thought – keeping calming thoughts in mind. You can use those as well, but then they are more like helpful thoughts.* • *A more realistic thought is a combination of your catastrophic thought and common sense: If I start feeling dizzy, and start having palpitations, that means I'm starting to have a hyperventilation or panic attack, and not that I'll get into an accident. Even though I think I'm losing control, this doesn't mean it will actually happen (the chance of me having or causing an accident is smaller than I think).*
	Guiding patients through CBT assignments (including exposure), for example by asking questions or providing information or suggestions	• *Is it correct that the believability of the catastrophic thought becomes weaker once the experiment has ended? So do you believe less strongly that you are becoming unwell?* • *I think that, based on what you told me earlier, perhaps we could think of other exercises. It's not so much about all the things you're afraid of doing, but about things you do differently because of your fear. That could be always making sure there is someone nearby that could help you. You can see that in the way you formulate your answers to the exercises: it is especially difficult to do things on your own.* • *The moment that you flee after all, or start measuring before the tension has gone down – that is no longer exposure, and it won't work either.*

		Moreover, the anxiety will just increase. One solution, I think, is to think of achievable steps and of using things that can aid you, in such a way that you can follow through. • *In the evaluation of your experiment you write that you tend to overestimate negative outcomes, so that situations can be scary beforehand. What can you do to reduce this fear? What exercises from the treatment can you use to do this?* • *Of course, it could happen that someone does get angry when you don't know something or say something wrong. How would you feel about that? Could you handle it? What could help you with that?*
Scheduling appointment for f2f session		• *We'll see each other on Thursday 26th November at 13:00, see you then!*
		Examples from feedback messages non-adherent to instructions
		• *We already discussed this session during our face-to-face session.* • *I will open the next session for you.* • *We will discuss the assignments when we see each other again.*

C3 Examples of adherence to blended instructions

Box 1. Example of adherence to blended instructions in f2f session, protocol component *Psychoeducation*

Therapist: *This treatment will last 15 weeks and face-to-face sessions will alternate with online sessions.*

(...)

Therapist: *Let's take a look at the online platform.*

Patient: *Okay.*

Therapist: *It's really easy, so you'll be fine. Here you see the sessions and your progress, so you know where to start when you log on again. Now we'll have a look at the introduction session together.*

Patient: *Yes, I can see that in Tasks.*

Therapist: *That's right, you can see the introduction sessions there. And here you see what the content of the session is.*

Patient: *I see.*

Box 2a. Example of adherence to blended instructions in f2f session, protocol component *Discussing Previous Online Session*

Therapist: *I gave you some feedback in the online session. You had made a list of exposure activities and described your catastrophic thought. You described what happens when you get into your car very clearly. You think: "Oh no, I will stop at the next petrol station."*

Patient: *That's right.*

Therapist: *I think those are reactions to your catastrophic thought. It's not that thought that makes you feel anxious and that makes you want to stop driving the car. Because ... what might happen if you keep on driving? What's the most catastrophic thing that could happen?*

Box 2b. Example of partial adherence to blended instructions in f2f session, protocol component *Discussing Previous Online Session*

Therapist: *First I want to briefly discuss the online session. How did that go?*

Patient: *I recognised some things, some of these exercises I've done before. And um… I found it difficult to describe my own situation.*

Therapist: *Yes.*

Patient: *That's still difficult for me. The rest of it was clear. When you start working on it, you start thinking about your situation. That's good, I think….*

Therapist: *I read it and I gave you feedback. How did you feel about that?*

Patient: *I found it supportive.*

Therapist: *Good.*

Box 3. Example of adherence to blended instructions in f2f session, protocol component *Preparing Upcoming Online Session*

Therapist: *This week you'll keep your panic diary, you can do that online. That means you describe every panic attack in your diary. For example, if you have trouble breathing again and that makes you feel anxious.*

(…)

Therapist: *I'll send you feedback on the online session next Monday in the morning. Is that okay?*

Patient: *Yes.*

Box 4a. Example of feedback message in online session adhering to *blended instructions*

Your answers in the exercises are very clear, nice! ***(Generic therapeutic feedback)***

In the first exercise you described the panic attack clearly, and you notice that it helps to formulate alternative thoughts such as: Thousands of people take the bus, I used to do that as well, and that is normal behaviour. What you're telling yourself is that it is not a dangerous situation, and that the sensations of fear are essentially unnecessary. Henceforth you'll see that the catastrophic thought "I am fainting" goes down from a believability of 80% to 25%, and that the alternative thought becomes more believable, and that because of this you will feel less anxious. ***(CBT-specific feedback)***

Very good that you practised the "head shaking". Just as you indicate at the end: this is an uncomfortable feeling. But the question is whether that also means that you are going to faint/die. You notice that the catastrophic thought goes down from a believability of 85% to 20%. It is not necessary to be anxious about those sensations.

Concerning the exposure list, it may be useful to look at an activity that you could perform daily. That would make it easier to practise (or more often) and to lower the threshold. That is important, because for the next face-to-face session you will perform three activities from your exposure list. Don't forget to keep track of those activities in your exposure diary. ***(Generic therapeutic feedback + CBT-specific feedback)***

We'll discuss how that went on Monday next week. Good luck and see you then! ***(Scheduling appointment for f2f session)***

Box 4b. Example of feedback message in online session adhering to *blended instructions*

Thank you for doing the exercises. Things are looking good! I will once again give you some tips for each exercise: ***(Generic therapeutic feedback)***

(4.1 Describe a panic attack). I see that you're getting better at describing that. Under the heading Thoughts, I'd like to know whether you have any more thoughts that arise. To bring automatic thoughts to the surface you can try to replay the event like a short movie and to ask yourself: what am I thinking now? Because what? And then?

About the catastrophic thought: does that thought describe the essence of your feelings? I ask this because you indicate that you are scared, angry and sad. What is the essence of your anxiety? For example: the thoughts never go away, and then....

For the alternative thought: nice, and a positive description! What effect does that thought have on you? Would the feeling or behaviour be different to when you think the catastrophic thought? ***(CBT-specific feedback)***

(4.2 Interoceptive exposure) Well done, you already did 2 exercises! I also see that the believability of the catastrophic thought was strongly reduced after the exercise. So it's not necessary to be afraid of physical sensations. ***(CBT-specific feedback)***

(4.3 Exposure list) Great that you already managed to fill in 10 activities. It's important that the activities give you the opportunity to investigate the catastrophic thought, for example, "The thoughts never end". If an activity is accompanied by a different catastrophic thought, then it's good to describe it as well (if possible in the diary). ***(CBT-specific feedback)***

It's important to hold on to the alternative thought, and how you would react and feel, in the back of your mind. That way you can really challenge the thoughts. Example: Hyperventilating 1 minute –

Catastrophic thought: I'm going to faint. Feeling: fear. Behaviour: sitting down and breathing calmly.

Alternative thought: I have all these physical sensations, but do not faint. Feeling: fear. Behaviour: continuing and enduring the hyperventilation ***(Generic therapeutic feedback)***

For the next face-to-face session, you will perform three activities from your exposure list. Don't forget to keep track of those activities in your exposure diary. ***(Scheduling appointment for f2f session)***

We will discuss how things went on the 18th of May. Good luck and see you then!

Box 5a. Example of feedback message on a *repeated session*

I gave you feedback on the online session during the last f2f session. You can now proceed to the next online session.

Box 5b. Example of feedback message on a *repeated session*

We completed this session together, because you needed some help with it.

Box 5c. Example of feedback message on a *repeated session*

Thanks for completing this session, you can now proceed to the next session.

Acknowledgements

Acknowledgements

The PhD project described in these pages was completed in the Department (and later Section) of Clinical Psychology at VU University Amsterdam. I can genuinely say that anyone will be hard-pressed to find a more pleasant working environment. My departure from the university coincided with the first COVID-19 lockdown, and when I went to pick up my belongings, the hallways and offices were desolated. A surreal last day in work, that sadly left me without a chance to thank my many good colleagues in person.

Heleen, Jeroen, Tibor, Annet, and Jiska who supervised me; José, Arthur, and Marleen who helped me develop Moodbuster Lite; Ho-Ming, Ward, Koen, Mayke, and Pim who co-authored various papers in this thesis; Sander, Rosa, Claudi, Willlem-Paul, Mark, and Saskia who took the time to critically read and evaluate this thesis; Khadicha, Christiaan, Lisa, Niko, Leah, Joost, Eirini, Anne, Marketa, Chani, and Felix, from Team Riper; Johan, Sherida, Eugène, Resi, and Deborah from the secretary office; Tanja, Robin, Claudia, Gizem, Dilfa, Metta, Gitta, Cansu, Dave, Gwenda, Alyssa, Mandy, Clara, and Willeke, who at some point shared an office with me; the many other colleagues that I had whatever kind of nice time with, and anyone else that slipped my mind.

A big thanks to all of you, it has been a pleasure!

My time as a PhD student would have looked significantly different without the involvement of some of the people I hold dearest. You were there before I started this job, and well after it ended. You cheered me up, kept me going, distracted me, and listened to my complaints for many, many hours.

Mum, Dad, Tim, Jan, Iva, Geke, Angel, and last but definitely not least Sarah. I could dedicate another book to explaining why I am grateful to you. Who knows, maybe one day I will…

Thank you so much for your love and friendship, your support, and all the fun we have had. It means the world to me!