Internet-based treatment for eating disorders: bridging the treatment gap

Rohrbach, P.J.

Internet-Based Treatment for Eating Disorders: Bridging the Treatment Gap

Pieter Johannes Rohrbach

Internet-Based Treatment for Eating Disorders: Bridging the Treatment Gap
Pieter J. Rohrbach
PhD Dissertation

ISBN: 978-94-6458-831-6
Cover design: Nesse van der Meer & Pieter Rohrbach
Druk: Ridderprint | www.ridderprint.nl

The work described in this thesis was performed at GGZ Rivierduinen Eetstoornissen Ursula and the Leiden University Medical Center in Leiden, and was funded by ZonMw (grant number 636310001).

Internet-Based Treatment for Eating Disorders: Bridging the Treatment Gap

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op donderdag 19 januari 2023

klokke 13.45 uur

door

Pieter Johannes Rohrbach
geboren te Alphen aan den Rijn
in 1991

Promotores
Prof. dr. E. F. van Furth
Prof. dr. P. Spinhoven

Co-promotor
Dr. A. E. Dingemans

Promotiecommissie

Prof. dr. N. H. Chavannes	
Prof. dr. A. W. M. Evers	
Dr. L. Hakkaart - van Roijen	*Erasmus Universiteit Rotterdam*
Prof. dr. A. M. van Hemert	
Prof. dr. A. Jansen	*Universiteit Maastricht*

Contents

Chapter 1

Introduction

Introduction

A question that has sparked interest since the rapid progress in computer and internet technology, is whether the sector of mental health care can be improved using such technology. Indeed, especially since the COVID-19 pandemic, technological innovations have been creatively implemented in many mental health institutions to keep patient care on track. Examples are therapy through video conferencing and using online modules or smartphone applications to help patients work on recovery even when possibilities for direct contact are limited. However, to confidently use such innovations it is necessary to investigate if they work, under what conditions they work most effectively and how they compare to alternatives, such as regular face-to-face treatments. The question of how to optimally use technology in the mental health sector should therefore be broken down into smaller pieces that can individually be researched. In this thesis, eating disorders will be the topic of interest. Not only are these disorders burdensome on both an individual and societal level, but some distressing characteristics coinciding with eating disorders, such as high levels of shame and a reluctance to seek appropriate care, might be well targeted by technology-based interventions.

Eating Disorders

Eating disorders are mental disorders, indicating that they are characterized by dysfunctional behavioral or mental patterns (centered around eating) that cause considerable distress or impaired functioning. Eating disorders frequently lead to severe psychiatric and somatic complications, reductions in quality of life or even death (Smink et al., 2013). There is evidence to suggest that the lifetime prevalence of eating disorders (i.e., the proportion of the population that will experience an eating disorder at some point in their life) has been growing and is now around 7.8% (Galmiche et al., 2019). The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013) recognizes four distinct eating disorder diagnoses: binge eating disorder, bulimia nervosa, anorexia nervosa and other specified feeding and eating disorders. A short account on each disorder is presented in Table 1.

It must be noted that while differences between the diagnoses exist, there are many common factors, such as a negative self-image, over-evaluation of weight and shape and disturbances in regulating mood and impulsivity (Lampard et al., 2013). Moreover, over 70% of people with an eating disorder also experience other psychiatric difficulties, such as self-harming behaviors, anxiety and depressed mood (Keski-Rahkonen & Mustelin, 2016). These comorbid psychiatric problems as well as physical problems that coincide with the eating problems, such as chronic pain, diabetes or problems with weight, result in more frequent use of health care services of individuals with an eating disorder compared to those without (Ágh et al., 2016; Weissman & Rosselli, 2017). However, they often do not seek or get help specifically for their eating problems (Hart et al., 2011). Indeed, it may take up to 68 months before individuals with an eating disorder receive targeted help for their problems (Austin et al., 2021). This discrepancy between developing eating disorder symptoms and not receiving appropriate care is referred to as the treatment gap. Bridging this treatment

gap, or in other words getting people with an eating disorder into appropriate care more quickly, will go a long way in making eating disorder treatment more effective on a systemic level (Moessner & Bauer, 2017). But how do we go about bridging this treatment gap?

Table 1. Description of the four eating disorder diagnoses

Disorder	Description	Lifetime prevalence estimates (Galmiche et al., 2019)	Recovery rates after treatment (Eddy et al., 2017)
Binge eating disorder	Recurring episodes of eating an objectively large quantity of food (compared to most others in a similar situation) in a short period of time (within two hours) and simultaneously feeling out of control or unable to stop.	2.8% for women, 1.0% for men	Estimated between 19-77%
Bulimia nervosa	Frequent binge eating episodes that are compensated, for instance by self-induced vomiting, use of laxatives, fasting and excessive movement or exercise.	1.9% for women, 0.6% for men	Estimated between 55% and 68%
Anorexia nervosa	Restricted food intake (and weight loss) driven by an intense fear of gaining weight or being fat. The self-image of people with anorexia nervosa is distorted, so that they believe they are heavy even when they are (severely) underweight.	1.4% for women, 0.2% for men	63% recovery (after 22 year follow-up)
Other specified feeding and eating disorder	Diagnosed when behavioral or psychological patterns similar to that of other eating disorders are present, but symptoms do not fit one of the three other diagnoses.	4.3% for women, 3.6% for men	Unknown

eHealth for eating disorders

eHealth refers to the use of technologies such as the internet and smartphone applications in health care. Internet interventions can act as stand-alone intervention or as addition to face-to-face treatment, and they differ regarding their content and technical features. For example, they can be completely self-guided (no guidance, automated) or have some form of guidance. Many internet interventions contain email or telephone contact with a health professional (e.g., a psychologist) in addition to an automated part of the intervention, such as online modules with information and tasks a user can complete. Internet interventions

have been proposed as a possible solution for challenges in mental health care, such as long waiting lists, financial pressure and reaching underserved populations (Aardoom, Dingemans, & Van Furth, 2016). Indeed, eHealth might be especially useful to target eating disorders. For example, internet interventions can be made anonymous so that feelings of shame and unsafety might be more easily overcome (Bachner-Melman et al., 2018; McClay et al., 2014). Consequently, individuals with an eating disorder can arguably be reached more quickly through the internet than through gatekeepers such as the general practitioner. Successful implementation of internet interventions to improve access to psychological treatment has already been reached for other mental disorders, such as depression and anxiety (Titov et al., 2018). While implementation of online interventions for eating disorders in society is not yet common, multiple trials have been conducted to investigate the feasibility and effectiveness of internet interventions (Aardoom et al., 2013b; Beintner et al., 2012; Dölemeyer et al., 2013; Linardon et al., 2020; Loucas et al., 2014; Melioli et al., 2016; Pittock et al., 2018; Schlegl et al., 2015; Taylor et al., 2021; Wade & Wilksch, 2018; Zeiler et al., 2021a).Overall, internet interventions for eating disorders have been found to be effective in reducing eating disorder symptomatology, such as disordered eating behaviors, body dissatisfaction and fear of weight gain. An online self-help program that has been developed and researched in the Netherlands is Featback. It contains an automatic monitoring and feedback system to help users become aware of (the severity of) their eating problems and seek support, either from within their close surroundings or professionally. It has been found to be effective in reducing eating disorder related problems and feelings of anxiety and depression (Aardoom, Dingemans, Spinhoven, et al., 2016).

Guidance

Across mental disorders, there is evidence to suggest that incorporating guidance increases the effectiveness of internet-based interventions (Baumeister et al., 2014). However, for eating disorders, mixed results on the effectiveness of adding guidance were found (Yim & Schmidt, 2019b). For example, a meta-analysis investigating individual components of eHealth interventions for eating disorders, found that guidance did not moderate intervention effectiveness (Barakat et al., 2019). This is surprising given that online guidance is repeatedly found to be highly valued (Galmiche et al., 2019; Yim & Schmidt, 2019a). Accordingly, Aardoom, Dingemans, Spinhoven, et al. (2016) showed that while adding online chat or email support by a psychologist increased satisfaction with the Featback intervention, it did not result in increased effectiveness. An explanation for lacking effectiveness of psychologist support could be that individuals with an eating disorder still experienced barriers to fully engage with the offered support, such as shame, fear to be stigmatized and feeling misunderstood. These barriers may be lower when guidance is offered by experts by experience. Such expert patients have a lived experience of an eating disorder, are fully recovered and have been trained to use their own experience to help others who are currently struggling with eating problems. Expert patients might inspire hope of recovery because they are proof that recovery is possible (Simoni et al., 2011), and a shared background with current sufferers could help to bond more quickly compared to other health professionals (Montoya & Horton, 2012). There is some evidence to suggest that support from expert patients has beneficial

effects for both the individual with an eating disorder and the expert patients themselves (Fogarty et al., 2016). For example, when providing support expert patients are affirmed in what they have already accomplished, the challenges they have overcome and how well they are doing currently. For individuals receiving support from an expert patient, findings cautiously suggest that the support can enhance quality of life, relationships and adherence to an intervention (Cardi et al., 2019; McCarroll, 2012; Perez et al., 2014; Ramjan et al., 2017). In summary, expert-patient support might be a strategic alternative to support from other health professionals, especially in the context of low-threshold interventions (like Featback), aimed at individuals who do not yet receive fitting professional care. However, more high-quality research establishing its effectiveness is warranted.

Costs of eating disorders

The search for innovative interventions, such as internet interventions and expert-patient support, continues in an effort to improve the overall effectiveness of eating disorder treatment, but also to reduce treatment related costs. Indeed, the impact of mental illness on society is huge, ranking the top 3 of causes of global burden (Vigo et al., 2016). The worldwide economic costs of mental disorders was estimated to be 2.5 trillion US dollars in 2010 and might be as high as 6 trillion dollars by the year 2030 (Marquez & Saxena, 2016). These costs are considerable, but most countries spend a disproportionally small amount of their yearly health budget on mental health and fail to provide people with the mental health services they need (World Health Organization, 2021), warranting both policy changes and continued effort to find effective and inexpensive treatment options.

Looking more specifically at costs associated with eating disorders, health care costs for people with eating disorders were estimated to be 48% higher compared to people in the general population (Van Hoeken & Hoek, 2020). Expected annual health care costs for individuals with eating disorder were estimated to be between €888 and €55K (Ágh et al., 2016). According to a study in the US, the additional yearly costs of eating disorders might be as high as $12K per person with an eating disorder when also considering costs outside of health care, such as reduced work productivity and caregiver costs (Deloitte Access Economics, 2020). Consequently, timely, effective and inexpensive treatment might not only reduce the burden on the individual, but on society as a whole. E-mental health has been proposed as a cost-effective alternative to usual care (F. Griffiths et al., 2006), but economic evaluations are necessary to substantiate this (Hedman et al., 2012; Tate et al., 2009).

Economic evaluations in mental health care

Economic evaluations can help to determine whether the benefits of an intervention outweigh the costs associated with the intervention compared to an alternative (e.g., doing nothing, care as usual or another intervention). The goal is to inform on the extent to which treatment options are worth their costs, so that patients and society pay a fair price. Specifically, there is a scarcity of financial resources in the mental health care sector, so economic evaluations are important for clinicians and policy makers to aid decisions about which interventions or treatments should receive funds in order to provide the best possible care.

Utilities

Economic evaluations can take several forms, one of which is the cost-effectiveness analysis. With this method, costs and effects, such as the extent to which certain symptoms are reduced, of an intervention of interest are compared to the costs and effects of a different course of action. In a specific form of cost-effectiveness analyses (called cost-utility analyses), the effectiveness is operationalized by quality-adjusted life years (QALYs). QALYs help to determine a fair price for a treatment based on how well it works. Basically, the QALY is an established measure of all the benefits and detrimental effects of a certain course of action. For example, a cream for toenail fungus might have a small benefit for people, producing some QALYs. Possibly, an EMDR trauma treatment and a drug treatment for insomnia lead to more significant benefits such as enhanced quality of life or even life prolongation, producing more QALYs. However, the drug treatment for insomnia has some bad side-effects, reducing its QALYs. Consequently, courses of action across disorders can be compared in how much QALYs they produce, which can inform resource allocation decisions (e.g., spending more money on EMDR trauma treatment and less on toenail fungus cream). To calculate QALYs of an intervention, a group of people receiving an intervention are inquired after their appreciation of their life quality, referred to as utility, over a period of time during and after the intervention. Utility can vary from 0 (death) to 1 (perfect life quality). QALYs produced by an intervention, then, equals the utility value multiplied by the years lived with this utility.

Generally, utility values are derived from quality-of-life scores obtained from generic health questionnaires such as the EQ-5D (EuroQol Group, 1990) or Short-Form-Six-Dimensions (Brazier et al., 1998). Such questionnaires approach quality of life as one's perception of their position in relation to their life goals, which is affected by one's physical health, psychological state, level of independence, social relationships, and environmental factors (WHOQOL Group, 1995). Therefore, they assess different domains of functioning that are found to be important contributors to one's quality of life, like physical functioning, vitality, pain, social participation, and mental health. Nevertheless, although the use of utilities (and corresponding QALYs) is widespread in economic evaluations, generic health questionnaires are criticized for not capturing all relevant domains of quality of life (Coast, 2004; Pietersma et al., 2013). Certain aspects of quality of life that fall beyond (physical) health might be underestimated, such as experienced social support, psychological resilience and the capability to cope with illness. As a result, the effectiveness of an intervention might be underestimated in economic evaluations, especially for interventions outside of the traditional health care model, such as social care, public health, general well-being, chronic illness, elderly care and mental health (Goranitis et al., 2016; Mitchell et al., 2017).

One approach to deal with the criticism on generic health questionnaires is to assess quality of life in terms of capabilities (i.e., the extent to which someone is capable of doing what one wishes to do) instead of current functioning (i.e., what or how someone is actually doing in one's life). Several instruments to measure capabilities are used (Helter et al., 2020). In the Netherlands the ICECAP questionnaire (Al-Janabi et al., 2012) is recommended to be used alongside established generic health questionnaires when benefits of an intervention are expected that not only involve (physical) health, but a broader sense of well-being (Zorginstituut Nederland, 2015). While such a recommendation is understandable given the criticism on generic health questionnaires, capability instruments such as the ICECAP require

further examination to be used confidently. Ultimately, capability instruments might help to make economic evaluations in areas outside the traditional (somatic) health care context more valuable.

Costs

Economic analyses involve effectiveness measures, but also an assessment of costs. Costs that are considered depend on the perspective of the economic evaluation. A health care perspective includes intervention costs (e.g., personnel and used materials to execute an intervention) and health care costs, such as costs related to visits to a hospital or general practitioner. In the case of a societal perspective, non-health care costs, such as sick days from work, reduced productivity while at work and caregiver costs, are added to the health care costs. Generally, a societal perspective is preferred over narrower perspectives, as it gives a better understanding of the costs involved for all parties affected by the (medical) decisions that might follow from an economic evaluation (Fahkri et al., 2017).

Cost-effectiveness

We have established how, in the context of economic analyses, benefits (mostly QALYs) and costs (health care and non-health care) of treatments are assessed. One more ingredient is required to determine whether one course of action is cost-effective over an alternative. Indeed, in the easy case that an intervention is more effective and less costly compared to a different course of action, it is considered dominant. However, if it is more effective, but also more costly, it is harder to evaluate whether the intervention is considered cost-effective over the alternative: is the added benefit worth the extra costs? To answer this question, we require information on how much society values the benefits. This value is indicated by society's willingness to pay (WTP). In the Netherlands, acceptable WTP values for one QALY have been estimated to range between €20,000 for interventions in the context of 'low disease burden' to €80,000 in the context of severe diseases (Zwaap et al., 2015). With all three ingredients (i.e., intervention effectiveness, associated costs and WTP value per QALY), a cost-effectiveness analysis can be conducted to compare different courses of action. Cost-effectiveness analyses can complement effectiveness research, as they can guide decisions on how money should be distributed over various treatment options. Moreover, especially for innovative treatments that promise good effectiveness at low costs (e.g., internet interventions), cost-effectiveness research helps to determine the added value of such interventions to established treatment options.

Dissertation outline

The general aim of this dissertation is to investigate whether and how internet-based interventions are a valuable addition to the existing pallet of available treatment options for eating disorders. Ultimately, results obtained throughout this dissertation could help to make treatment options available for individuals with eating disorders who are currently not reached. Low threshold online interventions with appropriate guidance show promise in reaching this goal, but further corroboration is necessary (Aardoom, Dingemans, & Van Furth, 2016).

A first step is to demonstrate their effectiveness, to establish that individual patients are likely to experience a reduction in eating disorder symptoms or other important outcomes when participating in such internet-based interventions. However, effectiveness results do not tell the full story. Considering their cost-effectiveness compared to different courses of action is also important to help policy makers with resource allocation decisions and to better understand their impact when implemented in real-world settings.

In order to reach this aim, first, the effectiveness of two internet-based interventions and their combination for eating disorders in the Netherlands were investigated: 'Featback', a fully automated monitoring and feedback system, and expert-patient support. Second, the evidence of cost-effectiveness of e-mental health in general (not only for eating disorders) compared to care as usual was reviewed to verify the often made claim that e-mental health brings good effectiveness at low costs. Additionally, possible improvements in cost-effectiveness research were considered. Specifically, the criticism on using QALYs in cost-effectiveness analyses of interventions outside the area of (physical) health was addressed, by exploring the ICECAP-A and preparing it for economic evaluations in the Netherlands. Finally, returning to the area of eating disorders, the cost-effectiveness of Featback, expert-patient support, their combination and a waiting list control condition were compared. The content of the various chapters in this dissertation is detailed here.

In *chapter 2*, a study protocol is presented to introduce the internet-based interventions Featback, expert-patient support and their combination. The chapter describes what the interventions entail, the design of the randomized controlled trial used to study its (cost-)effectiveness and the planning of data handling and analyses. Specifically, participants were randomly assigned to four conditions to compare (1) Featback (2) email or chat support by an expert patient, and (3) the combination of both with (4) care as usual for eating disorders. After an intervention period of eight weeks, participants were followed for a period of one year. In *chapter 3*, results on the effectiveness of the three active interventions are described. Findings include differences between the three active interventions and care as usual regarding changes in eating disorder symptoms, anxiety and depression, self-efficacy and experienced social support. The chapter also presents data on the satisfaction, intervention usage and whether the intervention stimulated people to seek (professional) help. The third chapter informs on whether and how internet-based interventions might incorporate online guidance in the form of chat and email support from an expert patient. Furthermore, it contributes to the understanding of the added value of Featback and expert-patient support to existing treatment options for eating disorders.

The number of studies on economic evaluations of internet-based interventions, covering many mental disorders, is rapidly growing, increasing the body of evidence to determine whether e-mental health interventions in general have a favorable balance between costs and effects. Therefore, in *chapter 4*, the existing evidence on cost-effectiveness of e-mental health interventions compared to usual care was systematically reviewed. Data from the reviewed studies were pooled together in a meta-analysis to capture the cost-effectiveness of internet-based interventions for various mental disorders compared to usual care. This is the first study to pool cost-effectiveness data in an aggregate-data meta-analysis in the area of psychiatry.

Chapter 5 presents research on validating the Dutch translation of the ICECAP-A instrument, a questionnaire that measures quality of life in terms of capabilities. This instrument might be especially useful in cost-effectiveness research on interventions that produce benefits beyond (physical) health, which includes e-mental health interventions such as Featback. Therefore, it is important to examine exactly what the ICECAP-A measures and whether it does so reliably.

Chapter 6 further prepares the ICECAP-A for economic evaluations in the Netherlands, by presenting the development of a tariff based on the Dutch general population. The ICECAP-A questionnaire assesses someone's capabilities in five areas, namely the extent to which someone is able to (1) feel settled and secure, (2) have love, (3) be independent, (4) achieve and progress, and (5) have enjoyment and pleasure. If someone indicates to improve in the area of enjoyment and someone else indicates to improve to a similar degree in the area of achievement, both individuals may not experience a similar increase in life quality. Developing an understanding of which capabilities people value more or less (i.e., contribute more or less to life quality) may lead to a more precise measurement of life quality when using the ICECAP-A. Such a mapping of the extent to which some capabilities are preferred over others are captured in a tariff. In other words, an ICECAP-A Dutch general population tariff is an operationalization of which capabilities are considered more or less important in the general Dutch population. It can be used to transpose all possible answers on the ICECAP-A questionnaire to better fit one's quality of life, putting more weight on items that were considered more important. These transposed ICECAP-A scores may vary between 0 (not at all capable to do what one wishes to do) and 1 (fully able to do what one wishes to do) and are referred to as capability values. Capability values can be compared to utility values, which are used when determining QALYs, and can be used in economic evaluations similar to QALYs.

The previous three chapters more broadly discuss cost-effectiveness of e-mental health interventions and possibilities for improvement of such economic evaluations. *Chapter 7* returns to eating disorders and compares the cost-effectiveness of Featback, expert-patient support, the combination of Featback and expert-patient support and care as usual. Data were captured alongside the randomized controlled trial described in chapter 2. The study gives insight in the produced benefits (in terms of quality of life), intervention costs, health care costs and societal costs involved with the investigated interventions. Consequently, the results help to inform clinicians and policy makers on whether implementing Featback and expert-patient support for eating disorders is worth the investment.

Finally, the results and implications of the previous seven chapters are summarized and discussed together in the *general discussion*. Study strengths and limitations are mentioned and future research directions and clinical implications are explored.

Chapter 2

A randomized controlled trial of an Internet-based intervention for eating disorders and the added value of expert-patient support: study protocol

Trial registration: Netherlands Trial Register, NTR7065. Registered on 7 June 2018.

Published as: Rohrbach, P. J., Dingemans, A. E., Spinhoven, P., Van den Akker-Van Marle, E., Van Ginkel, J. R., Fokkema, M., Moessner, M., Bauer, S., & Van Furth, E. F. (2019). A randomized controlled trial of an internet-based intervention for eating disorders and the added value of expert-patient support: Study protocol. *Trials, 20*, 1–17. https://doi.org/10.1186/s13063-019-3574-2

Abstract

Background: E-mental health has become increasingly popular in interventions for individuals with eating disorders (EDs). It has the potential to offer low-threshold interventions and guide individuals to the needed care more promptly. Featback is such an Internet-based intervention and consists of psychoeducation and a fully automated monitoring and feedback system. Preliminary findings suggest Featback to be (cost-)effective in reducing ED symptomatology. Additionally, e-mail or chat support by a psychologist did not enhance the effectiveness of Featback. Support by an expert patient (someone with a lived experience of an ED) might be more effective, since that person can effectively model healthy behavior and enhance self-efficacy in individuals struggling with an ED. The present study aims to replicate and build on earlier findings by further investigating the (cost-)effectiveness of Featback and the added value of expert-patient support.

Methods: The study will be a randomized controlled trial with a two-by-two factorial design with repeated measures. The four conditions will be (1) Featback, in which participants receive automated feedback on a short monitoring questionnaire weekly, (2) Featback with weekly e-mail or chat support from an expert patient, (3) weekly support from an expert patient, and (4) a waiting list. Participants who are 16 years or older and have at least mild self-reported ED symptoms receive a baseline measure. Subsequently, they are randomized to one of the four conditions for 8 weeks. Participants will be assessed again post-intervention and at 3, 6, 9, and 12 months follow-up. The primary outcome measure will be ED psychopathology. Secondary outcome measures are experienced social support, self-efficacy, symptoms of anxiety and depression, user satisfaction, intervention usage, and help-seeking attitudes and behaviors.

Discussion: The current study is the first to investigate e-mental health in combination with expert-patient support for EDs and will add to the optimization of the delivery of Internet-based interventions and expert-patient support.

Background

E-mental health

Comorbidity, relapse, chronicity, and mortality are common in eating disorders (EDs), which indicates the seriousness of these psychiatric disorders (Keel & Brown, 2010; Smink et al., 2013). Unfortunately, many individuals with EDs do not receive appropriate healthcare (Hart et al., 2011). A study in the Netherlands showed that it often takes many years to recognize that one is suffering from an ED and more than 4 years to seek treatment (De la Rie et al., 2006). Explanations for not receiving fitting care seem to range from geographical and financial reasons to fear of loss of control, fear of stigmatization, and feelings of shame (Becker et al., 2010; S. Griffiths et al., 2015; S. Griffiths et al., 2018). A simulation study by Moessner and Bauer (2017) suggested that we can most effectively help people with an ED, not by improving existing treatments or aftercare, but by guiding them to care more quickly and focusing on prevention. Additionally, the earlier patients with an ED receive proper treatment, the higher the chances are for full recovery (Aardoom, Dingemans, & Van Furth, 2016). Recently, e-mental health (i. e., offering care or treatment via technological means such as websites, teleconferences, and smartphone applications) has been proposed as a solution to bridge this treatment gap that exists for individuals with an ED. E-mental health has the potential to reduce barriers to seek help, since it can provide inexpensive, anonymous, and easily accessible interventions (Aardoom, Dingemans, & Van Furth, 2016). Consequently, such low-threshold interventions could help to improve early detection and intervention of ED problems and to promptly guide individuals to more intensive care if needed.

Nevertheless, research regarding the effects of e-mental health on ED pathology and help-seeking behavior is still scarce. Results of a recent meta-analytic review (Melioli et al., 2016) demonstrated that Internet-based programs, of which most relied on cognitive behavioral principles, successfully decreased ED-related symptoms such as body dissatisfaction, symptoms of bulimia nervosa, shape and weight concerns, dietary restriction, and negative affect, and increased self-esteem and self-efficacy. Two examples of Internet prevention interventions that have been proven effective in randomized controlled trials (RCTs) are Student Bodies (Beintner et al., 2012) and the Body Project (Stice et al., 2017). Student Bodies is a cognitive-behavioral Internet-based program, including psychoeducation, self-monitoring journals, behavioral exercises, and weekly assignments, aimed at improving eating- and body-related issues in people at risk for developing an ED. The Body Project appeals to the same group, but it employs a dissonance-based approach, by letting users critique the thin body ideal in written, verbal, and behavioral exercises. The strength of such interventions is their ability to reach an underserved population. However, they should not be seen as a replacement for face-to-face treatment, but rather as an addition to the stepped-care treatment of EDs (Aardoom, Dingemans, & Van Furth, 2016). Additionally, confidence in results from RCTs regarding e-mental health for EDs (covering a wide range of interventions but excluding studies in which the therapist was the primary means of delivering the intervention) is generally low (often because of the high risk of bias and inconsistency and the indirectness of and imprecision in outcomes), so more solid research regarding the form and content of such interventions is needed (Loucas et al., 2014).

Naturally, there are limitations to low-threshold Internet interventions aimed at prevention and early intervention of EDs in the extent to which they can respond to the personal situation of users, especially when compared with interventions in which intensive and direct contact with a professional is possible, such as blended care. Nevertheless, it is important for low-threshold Internet-based interventions not to employ a "one-size-fits-all" approach. Indeed, not everyone profits from or prefers the same content of treatment, and the Internet is a highly suitable medium to convey interventions in a flexible and interactive way. The Internet-based program "Featback" combines prevention and (early) intervention for individuals with ED symptoms. The program aims to make users aware of their eating-related and underlying problems. Users are encouraged to share their problems with their environment and, for more severe problems, to seek professional help. The program can be used anonymously, reducing the barrier to subscribe. It contains psychoeducation, a fully automated symptom monitoring and feedback system, and weekly chat or e-mail contact with a coach. A more detailed account of Featback is presented in the "Interventions" section of this article. Featback is based on ES[S]PRIT (Bauer et al., 2009), a program originally developed in Germany. Research on ES[S]PRIT suggested the intervention is both feasible (Bauer et al., 2009) and acceptable (Lindenberg & Kordy, 2015) and improves self-efficacy in young individuals with ED-related problems (Lindenberg et al., 2011) and enhances help-seeking behaviors (Kindermann et al., 2016; Moessner et al., 2016).

The current research group has performed a first RCT investigating the effectiveness and cost-effectiveness of Featback (Aardoom et al., 2013a). Featback was offered with or without chat, Skype, or e-mail support from a therapist, which resulted in four conditions: (1) Featback only, (2) Featback with weekly support from a therapist, (3) Featback with support from a therapist three times a week, and (4) a waiting list control condition when Featback was complemented with therapist support once or three times a week. It was found that Featback (with or without support) was more effective in reducing symptoms of bulimia nervosa ($d = -0.16$) and symptoms of anxiety and depression ($d = -0.31$) than the waiting list control (Aardoom, Dingemans, Spinhoven, et al., 2016). Contrary to our expectations, no difference in effectiveness between the active interventions was found. Regarding the cost-effectiveness, it was found that Featback with or without therapist support represented good value for the money when compared to a waiting list (Aardoom, Dingemans, van Ginkel, et al., 2016), indicating that Featback might be a good alternative to care as usual, especially for individuals who experience difficulties in seeking professional help. Although no added effect on ED psychopathology was found, Featback users were significantly more satisfied with the intervention when Featback was complemented with weekly or three-weekly therapist support. Finally, moderator analyses showed that Featback was most effective for individuals with mild to moderate bulimia nervosa symptoms (Aardoom et al., 2017). The present study aims to follow up and build on these findings by further investigating the (cost-)effectiveness of Featback and by investigating the added value of expert-patient support.

Expert-patient support

An explanation of why additional support from a psychologist did not add to the effectiveness of Featback (Aardoom, Dingemans, Spinhoven, et al., 2016) may be that although individuals suffering from ED symptoms appreciate the empathy and support of therapists,

it may not be enough to reduce ED psychopathology. Support by expert patients (i. e., recovered individuals with a lived ED experience, also referred to as peers or mentors) may prove to be more effective for those reluctant to seek help and in the aftercare for individuals who have completed treatment and are at risk for relapse (Simoni et al., 2011). Specifically, expert patients may be more effective in changing behavior and inspiring hope of recovery, because of a perceived similarity and credibility (Simoni et al., 2011). The self-evident credibility of expert patients may make their interventions more valuable, since reliable (Latimer et al., 2010) and personalized (Covey, 2014) messages are found to be more effective in changing behavior. Additionally, the shared experiences and accompanying (perceived) similarity enhances experienced social support and feelings of closeness (Byrne, 1961; Curry & Dunbar, 2013; Launay & Dunbar, 2015; Montoya & Horton, 2012; Montoya et al., 2008) and various bonding behaviors (Dahlander & McFarland, 2013; Vass et al., 2018). The idea that people who share a common background or problem have a unique resource to offer each other appears to be at the heart of peer support (Hughes et al., 2009). Relatedly, expert patients are thought to be effective in enhancing self-efficacy in patients, since they can powerfully model health behaviors and enhance patients' belief in their own capabilities (Bandura, 1988; Dennis, 2003; Simoni et al., 2011), which is one of the primary goals of Featback. Concordantly, in the current study it is hypothesized that the credibility and the shared background of an expert patient and participant are sufficient to establish feelings of closeness and make participants more receptive to interventions aimed at reducing ED symptomatology and enhancing experienced social support and self-efficacy.

The body of literature on the effectiveness of support by expert patients is growing. Many studies have investigated expert-patient support for patients with chronic somatic illnesses in comparison to treatment as usual and found positive, albeit small, effects on self-efficacy, self-management, illness-related quality of life, and worry about the illness (Foster et al., 2007; C. J. Griffiths et al., 2005; Hughes et al., 2009; Kew et al., 2017; Mehlsen et al., 2017; Riddell et al., 2016; Rogers et al., 2008). Adding expert-patient support to usual care has also been found cost-effective for patients with various chronic somatic illnesses (Richardson et al., 2008). However, results on the effectiveness of expert-patient support in mental illness are mixed. For example, in some studies expert-patient support was associated with reductions in depressive symptomatology in patients with major depression and was found to be as effective as professionally administered treatment and superior to a waiting list (Bryan & Arkowitz, 2015; Pfeiffer et al., 2011). Furthermore, increases in self-efficacy were found in patients with severe mental illness who received expert-patient support in addition to treatment as usual in comparison to patients who received treatment as usual only (Mahlke et al., 2016). On the other hand, several studies found that adding expert-patient support to treatment as usual had no significant effect on psychopathology, quality of life, empowerment, or user satisfaction (Lloyd-Evans et al., 2014). In addition, the type and objectives of the expert patient interventions are highly heterogeneous (Campos et al., 2014), and confidence in both positive and null findings is repeatedly low because of the high risk of bias (Lloyd-Evans et al., 2014). This complicates assessment of the value of expert-patient support for mental illness and warrants further research. Regarding EDs, currently only a few studies have been completed (Fogarty et al., 2016). Perez et al. (2014) report that individuals recovering from an ED who are assigned to an expert patient indicate better relationships, a higher quality of life, and increased intervention usage than recovering individuals who are not assigned to an

expert patient. Results of two pilot studies are in line with these findings (McCarroll, 2012; Ramjan et al., 2017). However, there were no active control conditions in these studies. Additionally, Cardi et al. (2015) and Beveridge et al. (2018) are currently conducting trials on the topic of expert-patient support and EDs. In summary, expert patients may have positive effects on self-efficacy, belonging, and psychopathology (Fogarty et al., 2016), but findings are currently too circumstantial to provide convincing proof for the effectiveness of expert-patient support for EDs or recommendations on its implementation, so further investigation is necessary.

Aims and research questions

The current study builds on the study by Aardoom, Dingemans, Spinhoven, et al. (2016) by investigating whether the results regarding the effectiveness of Featback will hold. More specifically, the first aim is to investigate the (cost-)effectiveness of the Internet-based intervention Featback in comparison to Featback with support from an expert patient, support from an expert patient without Featback, and a waiting list control condition (WLC). The primary outcome measures of the current study are ED-related attitudes. Secondary outcome measures include self-efficacy, social support, symptoms of depression and anxiety, motivation to change, user satisfaction, intervention usage, and help-seeking attitudes and behaviors. Finally, cost-effectiveness will be evaluated through the reported quality of life, outcomes for patients in terms of capabilities, and medical and societal costs. We have two hypotheses accompanying the first research aim.

(H1) Our primary hypothesis is that Featback without expert-patient support, Featback with expert-patient support, and expert-patient support without Featback will be more effective in reducing ED psychopathology and more cost-effective compared to a waiting list.

(H2) Secondly, we hypothesize that the combination of expert-patient support plus an online intervention will be more effective in reducing ED psychopathology and more cost-effective compared to expert-patient support or Featback only and that the improved effectiveness will be maintained up until a year later.

The second aim of this study is to investigate predictors and moderators of intervention response to explore what works for whom. Predictors and moderators that will be tested as predictors or moderators of treatment response and/or intervention usage are age, gender, and educational level, motivation to change, social support, severity of ED symptoms, severity of symptoms of depression and anxiety, self-efficacy, self-esteem, and closeness or perceived similarity of participants with expert patients. Additionally, self-efficacy is examined as a mediator. Besides the exploratory tests, there are two hypotheses concerning the second aim of this study.

(H3) Since a perceived similarity of participants to expert patients might enhance the receptivity and self-efficacy of participants, it is hypothesized that participants who feel more similar to the expert patient they are assigned to have better outcomes in terms of ED symptomatology, self-efficacy, and experienced social support.

(H4) It is hypothesized that, since the effectiveness of expert patients is theorized to come from effectively improving self-efficacy, changes in self-efficacy during the intervention period mediate subsequent long-term effects of the intervention on ED-related symptoms and experienced social support.

Thirdly, we aim to investigate practical experiences with Featback, such as intervention usage and user satisfaction.

(H5) It is expected that Featback with weekly expert-patient support will enhance intervention usage as well as satisfaction with the intervention compared to Featback alone.

Methods

Design

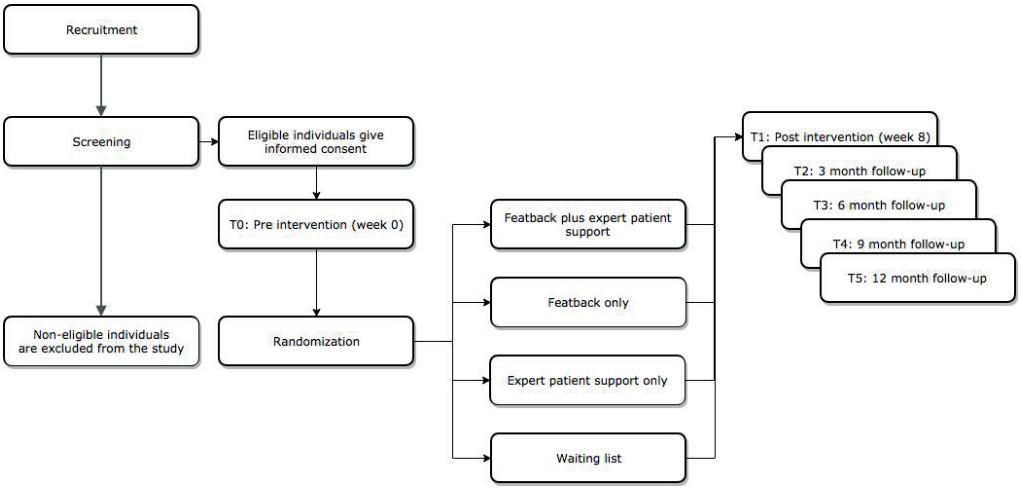
Since the present study is a continuation of previous work of this research group, the methodology described here will be similar and in some parts identical to the previous design (Aardoom et al., 2013a). The current study describes an RCT with a two-by-two factorial design with repeated measures to create four different conditions: (1) Featback, comprising psychoeducation and a fully automated self-monitoring and feedback system, (2) Featback with weekly individualized support by an expert patient through e-mail or chat, (3) weekly individualized support by an expert patient through e-mail or chat, and (4) a waiting list. A description of the content of the interventions for each condition is presented below. After screening, all eligible participants are asked to give informed consent and fill in online baseline measures (T0). Subsequently, they will be randomized to one of the four conditions. An independent researcher will conduct randomized allocation by using the SPSS function to produce random numbers. Hence, the main researcher will be blind to the randomization process. Randomization will take place in blocks of 40 participants. The current design does not allow expert patients to be blinded to the study goal, since they are required to help participants with their ED or ED-related problems to the best of their abilities within the intervention protocol. Naturally, expert patients know that the individuals they have contact with via e-mail or chat are randomized to one of the expert-patient support conditions. Similarly, participants are not blinded concerning the condition allocation.

After 8 weeks (intervention period or waiting period), participants are invited to complete the post-intervention assessment (T1). Finally, a link to the online follow-up questionnaires will be sent to them 3 (T2), 6 (T3), 9 (T4), and 12 (T5) months after T1 (see Figure 1). Ethical approval has been obtained by an independent medical ethics committee (CME LUMC Leiden, file number NL64553.058.18). The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) checklist is provided in Appendix A.1.

Participants

The study sample will be recruited via the Dutch e-community "Proud2Bme" (http://www.proud2bme.nl), via the Featback website and the network of the patient organization WEET. Proud2Bme is an interactive website that is designed for young people or adolescents (mainly

Figure 1. Flowchart of study procedures



girls) with eating problems or an ED. It is a healthy alternative to many pro-anorexia websites and promotes a healthy lifestyle and positive self-image. Eligible participants are aged 16 years or older, have access to the Internet, have self-reported ED symptoms defined as scoring 52 or higher on the Weight Concerns Scale (WCS) (Killen et al., 1993), or report one or more of the following ED symptoms assessed by the Short Evaluation of Eating Disorders (SEED) (Bauer et al., 2005): a body mass index (BMI) lower than or equal to 18.5, one or more binge eating episodes a week over the past 4 weeks, or one or more compensatory behaviors a week over the past 4 weeks. Participants are excluded if they are younger than 16 years or do not report any ED symptoms. Otherwise, there are no exclusion criteria, since both people with beginning and severe eating problems may benefit from Featback and/or expert-patient support (see also the "Ethical considerations" section of this article).

Interventions

Featback

All participants in the Featback conditions can access the Featback website on which comprehensive information on EDs and their causes and consequences can be found (i.e., psychoeducation). The psychoeducation will be purely self-guided, meaning that participants are free to choose what to read and when. For the monitoring and feedback system, participants receive an invitation by e-mail to complete a monitoring assessment weekly. This questionnaire consists of four 4-point Likert items assessing ED-related behaviors, namely (1) excessive concerns with body weight and shape, (2) unbalanced nutrition and dieting, (3) binge eating, and (4) compensatory behaviors. When participants have completed the questionnaire, a supportive feedback message will be automatically generated according to a pre-defined algorithm, which addresses their reported behaviors (healthy or unhealthy range)

and patterns of change (improved, deteriorated, or unchanged) on each of the assessed ED-related behaviors. Hence, the automated messages vary in content depending on the problems that users report. The messages contain a summary of self-reported eating problems, psychoeducation, and guidance on how to counter ED-related symptoms, which are formulated in a supportive and reinforcing way. Table 1 illustrates one of these automatically generated messages.

Table 1. Example of an automatically generated Featback message

"We are concerned about your eating problems. You indicate that last week you have been worrying about food and your body, you have not eaten sufficiently every day and have had more days on which you dealt with binges and compensatory behaviors. That's no small thing you're dealing with :(.
The urge to eat can emerge from stress, tension and/or emotions that suddenly occur. Is that something you recognize? Do one or more of these factors also precede a binge for you? It is possible to directly respond to these tensions or emotions by giving in to your binge. However, in fact you are not really heeding them, but you are muffling or dampening them and putting them aside. This mostly has a reversed effect, since not only do these tensions and emotions return at a later time, you generally feel worse after a binge as well.
Next time you feel an urge to binge or compensate your food, try to delay it. You will notice that after a while the binge or compensating behavior seems less necessary, or even not necessary at all! For this week, try to delay the urge for about 10 minutes. Also think about activities you can undertake during those 10 minutes to make delaying your binge or compensating behavior more bearable. Call a friend, put on your favorite music, go on a stroll through town or find another activity. Did you achieve the 10-minute delay? Excellent! Challenge yourself to extent the time you set for yourself every now and then.
Will you rise to the challenge? We are very curious to see what will happen when you learn to delay your harmful eating behaviors and whether this will help you. Good luck!"

Note. This message reflects the content when a participant indicates in the weekly monitoring that she or he worries about food, has at least one day in which she or he restricts food intake, and has had more binges and compensation behavior than in the previous week. The message is translated freely from the Dutch version

Weekly expert-patient support

All recruited expert patients ($N = 5$) have participated in an intensive day of training. The first part of the training comprised how to use the experience of having had and overcoming an ED to help others struggling with ED-related problems. The second part comprised an elaborate explanation of the research and the Featback program. Subsequently, a training specifically focused on the delivery of online support via chat and e-mail was delivered. An intervention protocol was handed out and explained to the expert-patient support team. The protocol includes guidelines about how to provide support so that all expert patients will work from a similar perspective and with similar methods. The five-phase model on which the intervention is based contains (1) warm welcome, (2) clarify the question, (3) determine the goal of the conversation, (4) elaborate on the goal of the conversation, and (5) close the circle. The phases of e-mail support are (1) extract the question, (2) formulate an answer, and (3) check and send the message. More detailed information on the models for e-mail

and chat support can be found in the handbook written by Schalken et al. (2010). The expert-patient supporters have practiced with offering chat and email support during and after the training, and feedback on their practice sessions was provided by an expert patient and experienced psychologist. Participating expert patients will receive monthly supervision during the study by an experienced expert patient and clinical psychologist to ensure high-quality and ethically correct support. They have a set amount of hours per week that they can flexibly distribute, and they receive monthly payment for worked hours on the project.

Participants can schedule a weekly appointment with an expert patient. For each session, participants can choose to receive support via chat or e-mail. Chat sessions have a duration of 20 min, and for e-mail support participants are required to send an e-mail before the scheduled appointment to which an expert patient will reply at the time of the appointment.

Waiting list control

Participants will be placed on a waiting list for 14 months (matching the participation duration of participants in the other conditions; 8-week intervention period plus 1 year follow-up), after which they will be offered 8 weeks of Featback with support from an expert patient. Participants in this condition will be asked to complete the same assessements as participants in the other conditions (i.e., T0-T5). Note that participants in all conditions are allowed to seek and receive treatment and take medication.

Intervention check

To assess the difference between support by expert patients and psychologists, a formal integrity check will be conducted. After data collection, 15 randomly selected chat and 15 e-mail sessions of expert patients will be compared to 15 randomly selected chat and 15 e-mail sessions of psychologists respectively (taken from the previous RCT of this research group; 20). Subsequently, three independent master level psychology students, blind to the source of the e-mail or chat session, will rate the 60 sessions with the integrity list. Expert patients and psychologists collaborated to create the integrity checklist, which involves (1) the structure of the session, (2) the content/interventions used during the session, and (3) the way in which these interventions were conveyed (see Appendix A.2. We expect that the structure of an e-mail or chat is similar between expert patients and psychologists, since the same structuring methodology is used. However, psychologists are expected to use a broader pallet of interventions (i.e., more distinct interventions) during the e-mail and chat support. Lastly, the most noticeable difference is expected in the way in which interventions are conveyed. More specifically, it is hypothesized that expert patients will explicitly mention their own experiences during every email or chat session to try to change attitudes or behaviors of participants, whereas psychologists will never do this. Additionally, expert patients are expected to use fewer medical terms or abbreviations in their e-mail or chat sessions than psychologists.

Measures

Table 2 presents an overview of the assessment instruments used for each measurement time. Estimated times (minutes) to complete each questionnaire are presented in parentheses. Details of the instruments are described in the following sections.

Screening measures

Weight Concerns Scale The Weight Concerns Scale (WCS) (Killen et al., 1993) is a five-item questionnaire used to evaluate the eligibility of participants. The five items are derived from a principal component analysis of a list of questions used to measure ED symptoms (Killen et al., 1993) and assess the extent to which participants struggle with their weight, eating pattern, shape, and perceived corpulence. Test-retest reliability and predictive validity have been investigated and demonstrated for the WCS (Killen et al., 1994). Furthermore, the WCS has been found to be able to predict students at risk for developing an ED (Jacobi et al., 2004).

Short Evaluation of Eating Disorders The Short Evaluation of Eating Disorders (SEED) (Bauer et al., 2005) contains six self-report questions designed to quickly assess the key ED symptoms. Participants are asked to evaluate their own body on several dimensions (e.g., thinness, attractiveness, and muscularity) and to report the frequency of several ED-related behaviors, such as self-induced vomiting, use of laxatives, and binge eating, over the last 4 weeks. Items are presented on a 5-point Likert scale. Summing the items of the two separate diagnoses leads to a severity index (range 0–3), with higher scores indicating higher severity. The SEED has been found to have good construct and criterion validity and was demonstrated to be sensitive to symptom change (Bauer et al., 2005).

Primary outcome measures

The primary outcome measures involve the Eating Disorder Examination Questionnaire (EDE-Q) (Fairburn & Beglin, 2008), which will be used to assess ED symptomatology. The EDE-Q has 28 items and assesses both the frequency of core ED behaviors (6 items) and ED-related attitudes (22 items) over the past 28 days. Items assessing the ED-related attitudes are presented on a 7-point Likert scale (range 0 "not at all" to 6 "every day/markedly") and include questions regarding weight, shape, and eating concerns and restraint. A global ED psychopathology score will be calculated by summing and averaging the 22 items. igher scores indicate higher ED psychopathology. Internal consistency, test-retest reliability, and discriminative validity of the EDE-Q have been found to be acceptable to high (Berg et al., 2011).

Secondary outcome measures

General Self-Efficacy Scale The General Self-Efficacy Scale (GSES) (Schwarzer & Jerusalem, 1995) is a 10-item psychometric scale designed to measure a general sense of perceived self-efficacy with the aim to predict coping with daily hassles and adaptation after experiencing various stressful life events. The questionnaire has been used in many studies with numerous

Table 2. Overview of assessment occasions and their content

		Study Period							
	Enrol-ment	**Base-line**	**Rando-mization**	**Inter-vention Period**	**Follow-up.** Post intervention, 3, 6, 9 and 12 month FU (T1-T5)				
Timepoint	-T1 (3)	T0 (23)		8 week in-tervention	T1 (37)	T2 (32)	T3 (32)	T4 (32)	T5 (32)
Enrolment:									
Eligibility screen	X								
Informed consent	X								
Randomization			X						
Interventions:									
Featback				[=====]					
Featback + Expert-patient support				[=====]					
Expert-patient support				[=====]					
Waiting list				[=====]					
Assessments:									
WCS (1)	X								
SEED (2)	X								
Demographics and Internet Usage (3)		X							
PHQ-4 (1)		X			X	X	X	X	X
EDE-Q (8)		X			X	X	X	X	X
Motivation to Change (1)		X							
GSES (3)		X			X	X	X	X	X
SSL (5)		X			X	X	X	X	X
RSES (1)		X							
User Satisfaction (5)					X				
Help-seeking attitudes and behaviors (5)					X	X	X	X	X
EQ-5D-5L (1)					X	X	X	X	X
ICECAP-A (1)					X	X	X	X	X
TiC-P MIDI (5)					X	X	X	X	X
PCQ (3)					X	X	X	X	X

Note: The Inclusion of Other in the Self scale (IOS scale; 0.5 min to complete) will be sent at week 3 of the intervention for all participants in a condition with expert-patient support. Attrition follow-up questions will be sent only to participants who do not respond to the assessments.
FU follow-up, *WCS* Weight Concerns Scale, *SEED* Short Evaluation of Eating Disorders, *PHQ-4* Patient Health Questionnaire, *EDE-Q* Eating Disorder Examination Questionnaire, *GSES* General Self-Efficacy Scale, *SSL* Social Support List, *RSES* Rosenberg Self-Esteem Scale, *EQ-5D-5 L* EuroQol five dimensions, five levels generic health index, *ICECAP-A* ICEpop CAPability measure for Adults, *TiC-P MIDI* Trimbos/iMTA questionnaire for Costs associated with Psychiatric Illness Midi version, *PCQ* Productivity Costs Questionnaire

participants (Schwarzer & Jerusalem, 2014). Responses are recorded on a 4-point scale. The individual items are summed to produce the final composite score with a range of 10 to 40. In samples from 25 nations (including the Netherlands), Cronbach's α ranged from 0.75 to 0.91, with most values over 0.80. Psychometric properties of the GSES are adequate (Luszczynska et al., 2005). The scale is one dimensional and designed for the general adult population, including adolescents.

Social Support List Perceived social support is measured with the short version of the Social Support List Interaction (SSL-12-I) (van Eijk et al., 1994). This self-report questionnaire measures the extent to which a participant experiences social support. The SSL-12-I contains 12 items in three scales, namely (1) everyday social support, (2) support in problem situations, and (3) esteem support (i.e., support resulting in self-esteem). Items (starting with "Does it ever happen that people..." and ending with statements like "...comfort you?" or "...give you good advice?") are presented on a 4-point scale ranging from "hardly or never" to "very often". Scores range from 12 to 48, and higher scores are indicative of more experienced social support. Psychometric properties are demonstrated to be good (Kempen & van Eijk, 1995; van Eijk et al., 1994).

Patient Health Questionnaire The Patient Health Questionnaire (PHQ-4) (Kroenke et al., 2009) measures symptoms of depression and anxiety. It consists of two primary anxiety items and two primary depression items. The anxiety and depression subscales have been found to reflect two separate dimensions (Kroenke et al., 2009). The four items are presented on a 4-point Likert scale ranging from 0 "not at all" to 3 "nearly every day". By summing all items, a composite score (range 0–12) can be calculated. Higher scores indicate higher pathology. The PHQ-4 has been demonstrated to possess factorial and construct validity (Kroenke et al., 2009).

Motivation to change Three items will be used to assess participants' motivation to change (Bewell & Carter, 2008; Genders & Tchanturia, 2010). The first item assesses the perceived importance to change of participants (*On a scale of 1 to 10, how important is it for you to change?*). The second item assesses the ability or confidence to change (*On a scale of 1 to 10, how confident are you that you could make a change if you wanted to?*). The third item assesses one's readiness to change (*On a scale of 1 to 10, how ready, or how prepared are you to change? Are you not prepared to change, already changing, or somewhere in the middle?*).

User satisfaction To assess the user satisfaction of Featback, a questionnaire was developed. Among other questions, participants are asked how they rate the quality of support they have received from Featback, whether Featback helped them to more effectively cope with their eating-related problems, and the extent to which they were satisfied with Featback in general. Additionally, participants are requested to rate the various components of Featback and address positive points as well as points for improvement (e.g., they are asked what they liked or disliked most and how the intervention can be developed further).

Help-seeking attitudes and behavior questionnaire A custom-made questionnaire was developed to assess help-seeking attitudes and behaviors. Participants are presented with

three 7-point Likert scale questions (ranging from 0 "not at all applicable" to 6 "fully applicable") concerning the extent to which they believe professional help is useful, they need professional help themselves, and they know where to find help. Furthermore, intention to seek professional help and actual help-seeking behavior are assessed, and the extent to which Featback contributed to these processes is inquired. Depending on whether the participant has sought help or not, the number of questions in this section ranges from three to five; they are either open or yes-no questions. Finally, participants are inquired as to the frequency of visiting websites other than Featback in relation to their (eating) problems, visiting and/or using a forum, or making use of online support service in relation to their (eating) problems.

Intervention usage Intervention usage will be operationalized by the amount of weekly monitoring assessments a participant has completed during the intervention period (range 0–8). Additionally, to be able to further investigate the relation between intervention usage and the effectiveness of Featback, the number of received support sessions of participants (one for each received e-mail or chat session) will be recorded.

Cost-effectiveness measures

General quality of life The EuroQol five dimensions, five levels generic health index (EQ-5D-5L) (EuroQol Group, 1990) is a standardized self-report questionnaire consisting of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). Participants rate each dimension on five levels (ranging from "no problems" to "extreme problems"). Consequently, 243 distinct health states are defined, each with a unique utility score, ranging from 1 (perfect health) to 0 (death). The health descriptions will be linked to empirical valuations of the Dutch general public, allowing utilities to be computed (Versteegh, Vermeulen, et al., 2016).

While the EQ-5D-5 L is the gold standard for computing utilities and economic evaluations, it is limited in the sense that it mainly addresses physical aspects of the health experience and might not be appropriate for assessing mental health problems. For example, psychiatric patients may not endure many physical problems while still experiencing considerable distress. Approaching outcomes for individuals by focusing on people's capabilities, instead of physical aspects of health, might therefore be more suitable for individuals with psychiatric conditions (Mitchell et al., 2017). The ICEpop CAPability measure for Adults (ICECAP-A) (Al-Janabi et al., 2012) shows promise in going beyond the general health status and capturing broader outcomes for individuals. The questionnaire aims to measure five capabilities on a 4-point scale, namely stability (the extent to which someone feels consistency and safety in life), attachment (the extent to which someone feels love, friendship, and support in life), autonomy (the extent to which someone can be independent in life), achievement (the extent to which someone can make progress in life), and enjoyment (the extent to which someone can enjoy life). The five items attempt to capture an individual's capability to live a life that he/she values. The ICECAP-A has been found to be a valid measure of one's capabilities (Keeley et al., 2013) and appears to be suitable for economic evaluation of outcomes of adults with mental health problems, such as depression (Mitchell et al., 2017).

Direct medical costs The Trimbos/iMTA questionnaire for Costs associated with Psychiatric Illness (TiC-P, Midi version) (Timman et al., 2015) will be used to calculate the total direct medical costs. The TiC-P assesses medical treatment utilization (e.g., number of contacts with the general practitioner, medical specialists, and paramedics) and medication use during the last 3 months. Additionally, the Midi version of the questionnaire is significantly shorter, reducing the burden for participants, while retaining 90% of the total cost estimated by the full version (Timman et al., 2015). The direct medical costs will be calculated using the Dutch guidelines for cost calculations in healthcare (Hakkaart-van Roijen et al., 2015). Reference unit prices of the corresponding health care services will be applied (Hakkaart-van Roijen et al., 2015; Kanters et al., 2017).

Demographics, closeness, self-esteem and attrition

Demographic A self-designed questionnaire will be used to gather demographic information. More specifically, gender, age, educational level, country of origin, and work situation will be assessed. Three additional questions are included, to inform about Internet access, the severity of eating problems, and whether participants have previously been or are currently in treatment for an ED.

Inclusion of Other in the Self scale The Inclusion of Other in the Self scale (IOS scale) (Aron et al., 1992) will be sent to participants allocated to one of the conditions with support by an expert patient at week 3 of the intervention to assess perceived similarity and feelings of closeness early in the working relationship between expert patient and participant. The IOS scale is a one-item instrument in which seven images with two circles representing the self and the other that overlap increasingly, from zero overlap to almost complete overlap, are presented to participants. The participants can then rate their relationship with another person by choosing the best fitting image. More overlap between the circles indicates a closer bond. This scale is particularly suited for the current study, since it can be used for any type of relationship, and not only romantic partners, friends, and acquaintances (Dibble et al., 2011). Furthermore, Gächter et al. (2015) conducted three studies to examine the validity and reliability of the IOS scale. A very strong convergent validity with a closeness index derived from six other relationship inventories (Spearman correlation of 0.85) was found. The authors conclude that the one-item IOS scale is easy to use, highly reliable, and a very powerful measure of closeness of relationships.

Rosenberg Self-Esteem Scale The Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965) is the most used measure of global self-esteem. It consists of 10 items measuring the affective evaluation of the self. The items are presented on a 4-point scale ranging from "totally agree" to "totally disagree". The questionnaire has been found to have satisfactory psychometric properties (Sinclair et al., 2010). The Dutch translation of the RSES originates from Franck et al. (2008), who created the translation using forward and back translation methods. Two studies indicate high internal consistency and good convergent and divergent validity of the Dutch version of the RSES (Everaert et al., 2010; Franck et al., 2008).

Attrition follow-up Attrition can refer to participants no longer using the intervention (non-usage attrition) or to participants completely dropping out of the study (dropout attrition).

Two questions were designed to investigate why people dropped out of the study or no longer used the intervention to which they were allocated.

If a participant fails to complete a monitoring assessment, an e-mail with a reminder will be sent. A week later, participants receive another reminder. This reminder includes a question asking whether one wishes to continue Featback or not, and, if not, participants are asked to answer one more question by writing down their reason(s) to quit the intervention (i.e., attrition follow-up question).

If a participant fails to complete a T1, T2, T3, T4, or T5 assessment within 1 week, an e-mail with a reminder will be sent. A week later, a second reminder will be sent. This reminder includes a question asking whether one wishes to further participate in our study, and, if not, participants are asked to answer one more question by writing down their reason(s) to quit the study.

Participant procedures

Figure 1 depicts an overview of the procedures that participants will undergo throughout the study. After recruitment, interested individuals can send an e-mail to the main researcher. Consequently, they will be sent a reply in which they are thanked for their interest and invited to complete a screening questionnaire. Participants who complete the screening questionnaire receive an email including feedback on their results and a notification about whether or not they are considered eligible for the study.

Participants who are eligible for the study will be sent an e-mail which contains an explanation of the study and corresponding procedures. Participants then have to give their informed consent to continue with the study. Agreeing to the terms of participation will be possible through clicking several checkboxes. Because informed consent is given online, the system will generate a popup at the end of the form with a question asking participants are sure they give consent to participate in the study. Here, participants will also be notified that they can leave the study at any time for any reason if they wish to do so, without any consequences. Subsequently, participants are presented with a link to the first assessment (T0). After giving informed consent and filling in the baseline measurement, participants will be randomized to one of the four conditions and will receive an email about the condition to which they are allocated and the corresponding procedures.

At this point, the 8-week period of intervention (or waiting) starts. At post-intervention (T1, week 8) and for all follow-up measurements (T2, 3 months; T3, 6 months; T4, 9 months; T5, 12 months) participants will be asked to complete the corresponding questionnaires. Finally, participants will receive an e-mail in which they are thanked for participating in the study. Participants who take part in the study receive 10 euros as compensation in the form of a gift voucher after the last measurement. Additionally, participating in the study will be made more personal and rewarding in the form of an e-mail sent to participants after every T1–T5 assessment, in which gratitude for participation is expressed. Hopefully, these incentives increase intervention usage and reduce attrition in the study.

Ethical considerations

Participants will be asked to complete questionnaires at baseline, a weekly monitoring questionnaire during the intervention, and a three-monthly questionnaire during the follow-up phase. This has been found to be an acceptable burden by a client panel. Earlier research with a comparable design showed no adverse effects of Featback (Aardoom, Dingemans, Spinhoven, et al., 2016), and participants in all conditions, including the waiting list, are allowed to seek treatment outside the study.

Individuals who enter the study and report severe ED symptoms during the screening or who develop severe ED symptoms during the intervention will be sent an email stating that their scores indicate serious ED problems and that professional help is warranted. More specifically, if a participant indicates a BMI of 15 or lower or reports compensatory behavior or bingeing at least every day during a week in the screening or Featback monitoring assessment, an alarm signal will be sent to the main researcher. If the participant is allocated to the Featback only or waiting list control condition, the participant will be sent an e-mail in which the researchers' concerns about the severe ED symptoms are expressed and recommendations for professional help are included. If the participant is allocated to one of the expert-patient support conditions and the participant has not scheduled an appointment for this week, an e-mail will be sent to encourage the participant to schedule an appointment. During the support session, the expert patient will discuss the alarm signal and severe ED symptoms and stimulate the participant to seek professional help. Similarly, participants in the support only condition who indicate (increasingly) severe problems in their chats or mails will be encouraged to seek professional help in subsequent sessions. If participants with severe ED symptoms do not make an appointment for a support session, the e-mail as described for participants in the Featback only or waiting list control condition will be sent.

Nevertheless, individuals with severe ED symptoms will not be excluded from the study, as there is no reason to withhold Featback or expert-patient support. It could well be that these individuals are reluctant to seek (face-to-face) treatment or that they are not fully aware of the severity of their symptoms. Accordingly, Featback may serve as an important first step to regular healthcare, because it could help individuals with the process of recognition and acknowledgement of the severity of their ED symptoms and the need to seek professional help (see also Moessner et al. 2016). Moreover, Featback and/or the individualized support from expert patients may serve as an important and unique source of support that could help individuals deal with their (eating) problems more effectively.

Expert patients are instructed to refer participants who report suicidal ideation to the website of "Stichting 113" (Stichting 113, 2009). The goal of this organization is to prevent suicide. They have psychologists, psychiatrists, and trained volunteers in employment who are accessible 24 h a day via telephone and chat.

Sample size calculation

An a priori statistical power analysis was conducted in G*Power version 3.1 to determine the optimal sample size having 80% power to detect a small effect size ($f = 0.15$, which corresponds to $d = 0.30$) between the active intervention conditions on the one hand and the control condition on the other. Consequently, the sample size calculation was based on

a between-factors repeated measures analysis of variance (ANOVA) with two groups and two measurements (i.e., baseline and post-intervention; T0–T1). A significance level of $\alpha = 0.05$ was maintained. The effect size was based on data from a previous RCT (Aardoom, Dingemans, Spinhoven, et al., 2016). Calculations indicated a total of 264 participants will be needed, meaning 88 participants per condition. We assume a Pearson correlation of 0.5 between the outcome variable on baseline and post-intervention (i.e., T0—T1), which explains 25% of the variance of the outcome variable. Therefore, the sample size per group can be reduced by 25%. However, adjusting for an anticipated dropout rate between T0 and T1 of 25% (based on previous data), we will still need 88 participants per condition ($N = 352$). The high dropout rate introduces a risk of bias through selective dropout. However, participants in the previous Featback trial (Aardoom, Dingemans, Spinhoven, et al., 2016) who dropped out during the intervention did not differ from those who did not drop out with regard to ED psychopathology, ED quality of life, comorbid anxiety or depression, age, weight, hours spent online, allocated condition, and duration of their eating problems. Hence, bias because of selective dropout in the current sample is, at least based on these variables, improbable.

Statistical analyses

All statistical analyses will be conducted in SPSS version 25 and the R statistical programming environment (R Core Team, 2017). A two-tailed significance level of $\alpha = .05$ will be maintained throughout the analyses unless indicated otherwise. All analyses will be conducted according to the intention-to-treat (ITT) approach. This means that all participants who underwent randomization, even those who withdrew from the study or deviated from the protocol, are included in the analyses. For the effectiveness analyses, both ITT and completers analyses will be conducted. A participant is considered a completer when he/she has completed at least five monitoring assessments (i.e., Featback only condition), five support sessions (i.e., expert-patient support only condition), or both (i.e., Featback plus expert-patient support condition). Missing data will be handled using multiple imputation (Rubin, 1987). Multiple imputations using predictive mean matching will be conducted in the statistical programming environment R (R Core Team, 2017). Interactions will be taken into account in the imputation procedure (Doove, Van Buuren, et al., 2014). Multiple imputation methods have several advantages over complete-case analyses or single imputation techniques and are therefore highly recommended (Rubin, 1987; Schafer & Graham, 2002). Homogeneity of groups (i.e., between-group differences) will be assessed at baseline (T0) using chi-square tests for categorical variables and ANOVAs for continuous variables. Additionally, homogeneity of variances will be assessed using Levene's statistic, and nonparametric testing will be used when appropriate.

Intervention effectiveness analyses

To investigate the effectiveness of Featback with and without weekly expert-patient support, within-group and between-group effect sizes (Cohen's d) will be calculated (i.e., the effects of time and intervention) using the pooled standard deviation of each group. We are mainly interested in the effect from baseline to post-intervention (T0-T1), but to see if effects are

maintained over the short and long term, analyses will be repeated for T1-T3 and T1-T5 respectively.

(H1) To answer the first and main hypothesis that the active interventions (i.e., Featback, expert-patient support, and Featback plus expert-patient support) are more effective than a waiting list in reducing ED symptomatology (i.e., primary outcome EDE-Q global score), the three active intervention conditions will be compared to the waiting list condition. Repeated measures ANOVA will be used for T0-T1 to test this hypothesis. To see if the effects are maintained, these analyses are repeated for T1-T3 and T1-T5.

(H2) The second hypothesis was that the combination of Featback and expert-patient support would be more effective than Featback or expert-patient support alone. A repeated measures ANOVA with post hoc tests will be conducted to compare T0-T1 differences in ED symptomatology between the four conditions. The post hoc analyses apply the Bonferroni correction for multiple testing. To see if these effects are maintained, these analyses are repeated for T1-T3 and T1-T5.

These confirmatory analyses will be repeated controlling for significant baseline variables (i.e., age, duration of ED psychopathology, number of psychological healthcare appointments). Additionally, the main analyses will be repeated for completers only.

Moderator and mediator analyses

Potential moderators of treatment effects will be investigated using model-based recursive partitioning methods (Hothorn & Zeileis, 2015). Model-based recursive partitioning can be used to detect what are called treatment-subgroup interactions. Treatment-subgroup interactions occur when subgroups of patients show differences in the effectiveness (i.e., a better or worse outcome) of one or more interventions. Model-based recursive partitioning can be used to identify these subgroups and their characteristics (Doove, Dusseldorp, et al., 2014; Fokkema et al., 2018) and ultimately help to tailor treatment to individual patients. These analyses will be explorative in nature and will apply the conservative Bonferroni correction for multiple testing.

(H3) The third hypothesis concerned the relationship between the closeness or perceived similarity of participants with the expert patient they are assigned to, and ED symptomatology, self-efficacy, and experienced social support respectively. The effects of perceived similarity on the three dependent variables will be investigated for the short term (i.e., gains until post-intervention, T1) and the long term (i.e., gains until the last followup, T5), resulting in six linear regression analyses that need to be conducted. We correct for multiple testing for these analyses using Holm's method.

(H4) To investigate the fourth hypothesis that changes in self-efficacy during the 8-week intervention period mediate subsequent changes in ED-related symptoms and experienced social support (and not vice versa), a cross-lagged panel design will be used. For self-efficacy and the two outcome variables, change scores will be calculated for

pre- (T0) to post-intervention (T1) and for post-intervention (T1) to long-term follow-up (T5; 12 months after T1). Next, hierarchical regressions will be performed, with post-intervention to long-term follow-up change of ED-related symptoms and experienced social support as dependent variables, and pre- to post-intervention change of self-efficacy as the independent variable. Additionally, the inverse relationships will be investigated. In other words, it will be examined whether changes in experienced social support or ED-related symptoms from pre- to post-intervention can predict post-intervention to long-term follow-up changes of self-efficacy. If the initial relationship is significant and the inverse relationship is not, a mediation effect of self-efficacy on ED-related symptoms and/or experienced social support is indicated. Corrections for autocorrelation and synchronous change will be applied to the mediation analysis.

Satisfaction and intervention usage analyses

(H5) To examine the fifth hypothesis that Featback with weekly expert-patient support results in an increased satisfaction and intervention usage, an ANOVA will be conducted to compare mean scores of satisfaction and intervention usage at T1 (i.e., directly after the intervention) between the three active intervention conditions. We correct for multiple testing for these analyses using Holm's method.

Cost-effectiveness analyses

The effects and costs of Featback and/or online support from an expert patient will be compared to those of usual care (i.e., participants in the waiting list condition) from a societal perspective, including healthcare costs and non-healthcare costs. No discounting will be applied due to the time horizon of the economic evaluation of 1 year.

The effects of an intervention will be assessed with the EQ-5D-5L at baseline and subsequent follow-up measurements. The EQ-5D-5L results will be translated into utilities using the EQ-5D-5L with Dutch rates (Versteegh, Knies, et al., 2016). The quality-adjusted life year (QALY) outcome per patient will be obtained by using the area-under-the-curve method for the utility scores obtained for each patient.

Outcomes of patients in terms of capabilities instead of QALYs will also be calculated using the ICECAP-A by means of the UK general population tariff (no Dutch tariff is available) (Flynn et al., 2015), since results from this instrument might reflect outcomes of psychiatric patients better than the EQ-5D-5L results (Mitchell et al., 2017).

The costs will be divided into healthcare costs and non-healthcare costs. Healthcare costs are calculated by summing the costs of Featback and/or online support from an expert patient, and other healthcare use during the first year of follow-up. Intervention costs of Featback include the maintenance of the program and website and payment to a psychologist following the procedure when a user develops severe ED pathology while using Featback. Intervention costs of expert-patient support are estimated by multiplying the time spent on sessions by their hourly pay rate. Costs for supervision of expert patients are also included, by multiplying the time spent at supervision by their hourly pay. Healthcare use is assessed with the TiC-P Midi (Timman et al., 2015), which records the number of contacts with care providers and use of medication during the last months (i.e., the period between every

follow-up measurement). The costs will be calculated by multiplying the number of contacts with a specific healthcare provider by the reference unit price of the corresponding healthcare service (Hakkaart-van Roijen et al., 2015; Kanters et al., 2017). Additionally, non-healthcare costs, including costs related to productivity losses at work through being absent or being less productive and having difficulties in performing unpaid work such as domestic tasks, are estimated. Costs related to absenteeism will be calculated according to the friction cost method, which means that the absent hours are multiplied by the average gross hourly wage per paid working individual in the Netherlands with a maximum of 12 weeks, the friction period in the Netherlands (Hakkaart-van Roijen et al., 2015; Kanters et al., 2017). The friction period is the timespan organizations need to restore the initial production level (Koopmanschap et al., 1995). Costs related to reduced efficiency at work are calculated based on the amount of hours of work participants estimate they need in order to catch up for all the work they were unable to perform because of health problems. These hours are again multiplied by the average gross hourly wage per paid working individual in the Netherlands, based on age and gender. Costs related to difficulties in performing unpaid work are calculated by multiplying the amount of hours that others would need to take over the unpaid work of the participant by the average gross hourly wage of a domestic worker. In summary, total costs of an intervention can be estimated by summing the healthcare costs (including the intervention costs) and non-healthcare costs consisting of productivity costs (including absenteeism, reduced efficiency at work, and difficulties performing unpaid work).

Differences in mean costs and effects per patient between interventions will be compared using a two-sided t test. The uncertainty regarding mean costs and effects per participant will be estimated using bootstrapping in Microsoft Excel, simulating 1000 bootstrap samples. Specifically, 1000 samples will be drawn from the original sample to estimate the sampling distribution and its 95% confidence interval. The results of the bootstrapping will be represented in cost-utility acceptability curves. These curves illustrate the probability that an intervention (i.e., Featback) is cost-effective in comparison with the alternative (i.e., care as usual) for a range of ceiling ratios, which are the maximum amount of costs a society is willing to pay for one unit change in outcome (i.e., QALY).

Data management

Data of participants will be handled and saved strictly confidentially according to the enforced laws and regulations, including the EU General Data Protection Regulation (GDPR) and the Declaration of Helsinki – 64th WMA General Assembly, Fortaleza, Brazil, October 2013. Data obtained from participants will be, among others, e-mail address, age, level of education, and data about the health of participants. Participants' e-mail addresses and other data that can be directly traced to them will be coded with a number so that their privacy is protected. Non-coded data will be saved separately, and only the main researchers, the accredited METC, and Inspectie Gezondheidszorg en Jeugd (IGJ) will have access to this data file. Data will be kept for a minimum of 10 years, according to guidelines from the Association of Universities in the Netherlands (VNSU). Participants can withdraw from the study at any moment without consequences. Data gathered from participants up until their withdrawal will still be used for analyses. No official data monitoring committee will be formed, since no difficulties in data management are anticipated, but use of a data log,

making back-ups of the anonymized data file, and regularly checking data completeness are several methods that will be employed to promote data quality.

Discussion

The aims of the current study are threefold. The first aim is to investigate the (cost-)effectiveness of the Internet-based self-help program Featback with and without expert-patient support. The second aim is to explore predictors, moderators, and mediators of intervention response, to better understand how and for whom Featback works. Thirdly, practical experiences with Featback, such as the intervention usage and user satisfaction, will be examined.

The current study design has several strengths. Firstly, it is highly similar to a previous RCT from the same research group. Therefore, findings regarding the effectiveness of Featback can be replicated, and limitations of the previous study can be overcome. Secondly, cost-effectiveness analyses of Internet-based interventions are rare but very useful in judging which interventions should be applied over others. Indeed, with ever-declining finances for health provisions and an increase in desire to offer effective and inexpensive treatment by health insurances, cost-effective analyses are indispensable. By conducting such an analysis, the current study aims to contribute to economically sensible choices regarding treatment for individuals with eating-related problems that are found to be effective as well. Thirdly, the ITT approach used for data analyses and the multiple imputations used for handling missing data are solid and recommended methods. Lastly, the relatively long follow-up period of 1 year helps to more fairly examine the effectiveness of the different interventions in the long term.

Additionally, some limitations of the current study design should be noted. Firstly, participants are allowed to engage in treatment outside of the research for ethical reasons. Although we will control for healthcare appointments during the analyses, methodologically it would be preferable to have all the participants only receive the experimental intervention. Lastly, only online measures will be completed by participants. This limits the diagnostic accuracy, introduces recall bias, and might reduce intervention usage as well. Indeed, we expect a fairly high dropout rate, and missing data might not be at random, which will need to be taken into account when analyzing the data. Nevertheless, the current approach is needed to maintain the low threshold and anonymity of the intervention, making it possible to generalize beyond this study to the real effects of Featback with and without expert-patient support.

The (cost-)effectiveness of Internet-based interventions in combination with expert-patient support for the (early) interventions of individuals with an ED or related symptoms has not been investigated before. Results on this subject will contribute to the delivery of e-mental health and expert-patient support and help to guide individuals with eating problems to the care they need.

Chapter 3

Effectiveness of an online self-help program, expert-patient support and their combination for eating disorders: results from a randomized controlled trial

Published as: Rohrbach, P. J., Dingemans, A. E., Spinhoven, P., Van Ginkel, J. R., Fokkema, M., Wilderjans, T. F., Bauer, S., & Van Furth, E. F. (2022). Effectiveness of an online self-help program, expert-patient support and their combination for eating disorders: results from a randomized controlled trial. *International Journal of Eating Disorders, 55*, 1361-1373, https://doi.org/10.1002/eat.23785

Abstract

Objective: Many individuals with an eating disorder do not receive appropriate care. Low-threshold interventions could help bridge this treatment gap. The study aim was to evaluate the effectiveness of Featback, a fully automated online self-help intervention, online expert-patient support and their combination.
Method: A randomized controlled trial with a 12-month follow-up period was conducted. Participants aged 16 or older with at least mild eating disorder symptoms were randomized to four conditions: (1) Featback, a fully automated online self-help intervention, (2) chat or email support from a recovered expert patient, (3) Featback with expert-patient support and (4) a waiting list control condition. The intervention period was 8 weeks and there was a total of six online assessments. The main outcome constituted reduction of eating disorder symptoms over time.
Results: Three hundred fifty five participants, of whom 43% had never received eating disorder treatment, were randomized. The three active interventions were superior to a waitlist in reducing eating disorder symptoms ($d = -0.38$), with no significant difference in effectiveness between the three interventions. Participants in conditions with expert-patient support were more satisfied with the intervention.
Discussion: Internet-based self-help, expert-patient support and their combination were effective in reducing eating disorder symptoms compared to a waiting list control condition. Guidance improved satisfaction with the internet intervention but not its effectiveness. Low-threshold interventions such as Featback and expert-patient support can reduce eating disorder symptoms and reach the large group of underserved individuals, complementing existing forms of eating disorder treatment.

Public significance statement

Individuals with eating-related problems who received (1) a fully automated internet-based intervention, (2) chat and e-mail support by a recovered individual or (3) their combination, experienced stronger reductions in eating disorder symptoms than those who received (4) usual care. Such brief and easy-access interventions play an important role in reaching individuals who are currently not reached by other forms of treatment.

Introduction

On average, it takes people with an eating disorder several years before they receive help specifically for their eating disorder (Austin et al., 2021). While individuals with an eating disorder generally utilize more health care services than those without, only a minority seeks targeted treatment (Hart et al., 2011; Weissman & Rosselli, 2017), indicating that many do not receive the care they need. This is worrying, since a longer duration of untreated eating disorder seems to be indicative of poorer outcome (Andrés-Pepiñá et al., 2020). The evident treatment gap (i.e., the discrepancy between people in need of help for their eating disorder and those actually receiving it) underscores the need to reach this underserved group, for example by using scalable, easily accessible, low-threshold interventions (Aardoom, Dingemans, & Van Furth, 2016; Kazdin et al., 2017; Moessner & Bauer, 2017). Two possible ways of making interventions more accessible for a large audience are internet interventions and support from an expert (recovered) patient.

Internet interventions for eating disorders

As captured in multiple systematic reviews and meta-analyses over the past decade (Aardoom et al., 2013b; Beintner et al., 2012; Dölemeyer et al., 2013; Linardon et al., 2020; Loucas et al., 2014; Melioli et al., 2016; Pittock et al., 2018; Schlegl et al., 2015; Wade & Wilksch, 2018; Zeiler et al., 2021b), internet interventions for eating disorders appear to have a beneficial effect on eating disorder symptoms and related complaints such as drive for thinness and weight and shape concerns compared to care as usual. Technological advancements have allowed internet-based interventions to be increasingly personalized towards users. The internet intervention 'Featback' is one such application that aims to provide low-threshold and easily accessible care for people with (symptoms of) an eating disorder, through personalized feedback. It is a brief online self-help program that works with an automated monitoring and feedback system. The main goal of Featback is to reduce eating disorder symptoms, by making users aware of their eating disorder symptoms, providing support and stimulating help-seeking behaviors, both towards their direct environment and professional facilities. An earlier randomized controlled trial showed Featback to be (cost-)effective in reducing symptoms of bulimia nervosa, depression and anxiety compared to a waiting list control condition (Aardoom, Dingemans, Spinhoven, et al., 2016; Aardoom, Dingemans, van Ginkel, et al., 2016). Interestingly, adding psychologist support once or three times a week increased satisfaction with the intervention, but not its (cost-)effectiveness.

The impression is that personal guidance adds to the effectiveness of, adherence to and

satisfaction with an online intervention, but this is mostly based on what we know from other disorders (Baumeister et al., 2014). Often, guidance is provided by therapists, but expert patients are increasingly involved in research and delivering support. Theoretically, expert-patient support can be especially valuable in providing low-threshold interventions, because a shared background and natural credibility enables them to establish a rapport, effectively model healthy behaviors and enhance self-efficacy (Dennis, 2003; Simoni et al., 2011). They might also be easier to approach and confide in than health professionals. Only few studies have involved expert-patient supporters (Fogarty et al., 2016; Lewis & Foye, 2022). Findings cautiously indicate that expert-patient support can enhance quality of life, relationships, and adherence to an intervention (Cardi et al., 2019; McCarroll, 2012; Perez et al., 2014; Ramjan et al., 2017). Concordantly, guidance from an expert patient might be a strategic alternative to support from health professionals. However, its added value and effective ways of implementation are not yet established.

Aims

The aim of the study was to investigate the effectiveness of Featback, expert-patient support and their combination. First, it was hypothesized that the three active interventions were more effective than a waiting list control condition. Secondly, the combination of Featback and expert-patient support was expected to be more effective than the two interventions separately. Lastly, no differences in effectiveness were anticipated between separately receiving Featback or expert-patient support.

Methods

This study was preregistered at the Dutch Trial Register (trialregister.nl/trials; identifier NL7065) and a study protocol with elaborate descriptions of the hypotheses and methods has been published (Rohrbach et al., 2019). The described repeated measures ANOVAs in the study protocol were altered to mixed model analyses, since they were more versatile. Additionally, a sensitivity analysis controlling for all relevant prognostic variables was performed rather than only with variables that differed significantly at baseline, as this is statistically preferable (De Boer et al., 2015). No other changes to the analysis plan per protocol were made. An economic evaluation concerning this trial has been published elsewhere (Rohrbach, Dingemans, Van Furth, et al., 2022). Results on prediction, moderation, and mediation will not be addressed here.

Design

A randomized controlled trial with a two-by-two factorial design with planned contrasts was used, creating four conditions: (1) Featback only, (2) Featback plus weekly expert-patient support, (3) weekly expert-patient support only, and (4) a waiting list control condition. Participants were assessed at baseline (T0), post intervention (eight weeks after baseline; T1) and 3 (T2), 6 (T3), 9 (T4) and 12 (T5) months after post intervention. All assessments consisted of online self-report questionnaires.

Participants and procedure

The majority of participants were recruited via Proud2Bme, a Dutch e-community for people with eating related problems. Other sources were also used, such as the Featback website, a blog on a fashion and health website aimed at (female) teenagers, social media, Google Ads and the Dutch eating disorder patient organization. Eligible participants were aged 16 years or older and had at least mild self-reported symptoms of an eating disorder; 52 or higher on the Weight Concerns Scale (Killen et al., 1993) or, as reported on the Short Evaluation of Eating Disorders (Bauer et al., 2005), a BMI of 18.5 or lower or at least weekly binge eating episodes or compensatory behaviors in the past four weeks. In the case of severe eating disorders, participants received the advice to seek professional help, but could still participate in the study as they may benefit from the offered interventions too. After expressing interest and filling out the screening questions, all eligible participants were asked to complete an online informed consent form and the baseline assessment (T0).

Participants were randomly allocated to one of the four conditions. An independent researcher created the allocation sequence using the SPSS function to produce random numbers. Hence, the randomization sequence, using blocks of 40 participants, was concealed from the principal investigator. The current design made it impossible to mask participants and expert patients to allocation. During the trial the principal investigator was not masked to the allocation, in order to send the appropriate information to participants. However, the content and timing of mails and reminders were standardized in order to avoid performance bias.

Interventions

Participants in all four conditions were allowed to seek help from any source for their eating disorder symptoms or other complaints. In this sense, the waiting list control condition can be regarded as treatment as usual for the current sample.

Featback

Participants in this condition received an account for the weekly monitoring and feedback system. During eight weeks participants received a weekly email with a link to a questionnaire with four questions on eating disorder related symptoms. On a 4-point scale, participants rated how often the behavior or symptom occurred this week, ranging from 'not at all', '1-3 days', '4-6 days' to '7 days'. The weekly feedback message was also dependent on whether participants indicated to have improved, deteriorated or stayed the same compared to previous week regarding the four monitoring questions. After completing the monitoring questionnaire, participants received a supportive feedback message that matched the participant's answers from a database with over 1250 different messages, written in collaboration with expert patients, scientists and psychologists. The supportive messages (on average 384 words) contain a summary of self-reported eating problems and changes compared to the previous week, psychoeducation, and guidance on how to counter eating disorder related symptoms. Additionally, participants could access the Featback website with psycho-educative material on eating disorders at their own convenience.

Expert-patient support

Expert patients ($N = 5$) received an intervention protocol and one day of training on how to use their own experience to help others with eating disorder symptoms via chat and email. During the trial, expert patients received monthly supervision from an experienced expert patient and clinical psychologist (EvF). Expert patients scheduled 20-minute slots flexibly across the week to give participants ample choice to access support. They worked 4 to 6 hours a week, could maximally support 12 participants at any one time and received monthly payment.

Participants in this condition were assigned to one of the expert patients for eight weeks and received an account to schedule appointments. For each of the eight sessions, participants were able to choose between email and chat support. Chat sessions closed automatically after 20 minutes. For email sessions participants were asked to send an email to their expert patient before the scheduled appointment, so that the expert patient could respond during the 20-minute appointment.

Featback with expert-patient support

Participants in this condition were able to make use of both Featback and weekly 20-minute chat or email support from an expert patient.

Waiting list control

Participants in this condition were placed on a waiting list for 14 months. The timing of the assessments was equal to that in the other conditions. After the waiting period, participants were offered eight weeks of Featback with weekly expert-patient support.

Intervention check

An intervention check, based on a checklist constructed before the trial started (Rohrbach et al., 2019), was performed to investigate whether and how sessions from expert patients were distinguishable from sessions by psychologists, as performed in the previous Featback trial by Aardoom, Dingemans, Spinhoven, et al. (2016). Two masked master-level psychology students rated the structure, intervention content and methods of delivery of 30 chat and email sessions from expert patients and 30 chat and email sessions from psychologists.

Outcomes

The primary outcome measure was eating disorder symptomatology as assessed by the Eating Disorder Examination Questionnaire (EDE-Q 6.0) (Fairburn & Beglin, 2008). A total score of eating disorder pathology was computed by taking the average of 22 items presented as 7-point Likert scale questions. Secondary outcomes measures included symptoms of anxiety and depression measured with the 4-item Patient Health Questionnaire (PHQ-4) (Kroenke et al., 2009), general self-efficacy measured with the General Self-Efficacy Scale (GSES) (Schwarzer & Jerusalem, 1995) and experienced social support measured with the 12-item Social Support List (SSL-12-I) (van Eijk et al., 1994). Additionally, motivation to

change, user satisfaction with the automated messages and expert-patient support, and help seeking intentions and behaviors were assessed with self-developed questionnaires. At baseline, self-esteem measured with the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965) was obtained, as well as demographic information including gender, age, educational level, country of origin, work situation, internet access, self-reported severity of eating problems and eating disorder treatment history. Lastly, the Inclusion of Other in the Self scale (IOS scale) (Aron et al., 1992) was used to assess, at week 3 of the intervention, the extent to which participants allocated to a condition with expert-patient support perceived themselves to be similar to the expert patient they were paired with. Psychometric properties of all questionnaires were adequate and can be found in the published protocol (Rohrbach et al., 2019).

Statistical procedures

All participants who underwent randomization were included in the analyses, following the intention-to-treat approach. An a priori sample size calculation, finding the optimal sample size for the main research question with a power of 80% to detect a small effect ($d = 0.30$), indicated 88 participants per condition ($N = 352$) were needed.

Main analyses were conducted in R version 4.0.2 (R Core Team, 2018) using linear mixed models (lmer function from the lme4 package) including random intercepts. To analyze the effect of the intervention type, three condition contrast were created. Specifically, to investigate the main hypothesis that the active interventions (i.e., Featback, expert-patient support and Featback plus expert-patient support) were more effective than a waiting list in decreasing eating disorder symptomatology, the three active conditions (pooled) were contrasted against the waiting list (CC1). This contrast allowed to investigate whether offering one of the active interventions, on average, resulted in greater symptoms reductions compared to the waiting list. The second contrast (CC2) distinguished the expert-patient support only and Featback only conditions (pooled) from the combination condition (i.e., Featback plus expert-patient support) to examine whether the combination condition was superior in reducing eating disorder symptoms. The third contrast (CC3) consisted of the Featback only versus expert-patient support only condition and informed on the relative effectiveness of offering only Featback and only expert-patient support. Moreover, five time contrasts were included, being baseline versus post intervention (TC1), and post intervention versus 3 (TC2), 6 (TC3), 9 (TC4) and 12 (TC5) month follow-up. Changes from baseline to post intervention (TC1) were of primary interest. The other time contrasts were used to inspect whether effects were maintained over time. All condition and time contrast combinations (15 in total) were tested separately to avoid noise in the models and improve interpretation. In other words, conditions or time points that were irrelevant for the effect of interest were removed from the model, as they may have introduced error variance, making the model less parsimonious. As recommended by (Cheng et al., 2010) when testing pre-specified models, Bonferroni adjustment of the p-values ($\alpha = .05/15$) was applied to account for multiple testing and reduce type I errors. In summary, CC1*TC1 was the interaction of primary interest. CC2*TC1 and CC3*TC1 were used to test our second and third hypotheses. To see the long-term effects of the interventions, TC5 (1-year follow-up) was deemed most informative. An overview of all statistical models can be found in Appendix B.1. These

main analyses were repeated for symptoms of anxiety and depression, general self-efficacy and experienced social support.

Additionally, six linear regression analyses were conducted to investigate both the short (T1) and the long term (T5) relationship between perceived similarity ratings, and (1) eating disorder symptomatology, (2) self-efficacy and (3) experienced social support. Multiple testing was accounted for using Holm's method (Holm, 1979).

Missing data

Missing data were multiply imputed (Rubin, 1987) using R. Logistic regression (multinomial) was used for imputing categorical variables, while predictive mean matching was used for most of the numerical variables (Rubin, 1986; Van Buuren, 2012). Variables constructed from other variables (e.g., BMI was determined by weight and length) were imputed using passive imputation (Van Buuren & Groothuis-Oudshoorn, 2011).

The number of predictors for each variable with missing data was determined by using a rule of thumb of 15 cases per predictor (Stevens, 2001). For a specific variable with missing data, the other variables that were most strongly related to this variable were chosen as predictors for the missing data. If the variable with missing data and a potential predictor were both numerical, then their absolute correlation was used as a measure of association. Partial η^2 was used as a measure of association if the variable with missing data was numerical and the potential predictor was categorical. Finally, if both the variable with missing data and a potential predictor were categorical, then Cramér's V was used. Missing data were imputed 100 times, creating 100 complete versions of the incomplete dataset.

Sensitivity analyses

The main analysis (CC1 and TC1) was repeated for participants with an adequate dose only. To be considered an adequate-dose participant, a participant in the Featback only, expert-patient support only and combination condition should have completed at least five out of eight monitoring assessments, five out of eight support sessions or both respectively. Secondly, the main analysis was repeated including covariates that were assumed to be prognostic for treatment outcome (i.e., age, baseline eating disorder symptoms, eating disorder duration, eating disorder treatment history, psychological health care visits, baseline self-esteem, baseline motivation to recover from the eating disorder, baseline anxiety and depression symptoms, baseline self-efficacy, baseline experienced social support and baseline BMI).

Results

Participants

Participant flow throughout the study is presented in Figure 1 and baseline characteristics are summarized in Table 1. Recruitment went as planned and lasted from October 2018 until October 2019 with the last follow-up measurement completed in December 2020. The baseline distributions of eating disorder symptoms across the sample are displayed in Figure

2. Study drop-out rates did not differ between conditions at post intervention, $\chi^2(3) = 3.99$, $p = .26$, or other assessments. Most mentioned reasons for dropping out, based on 26 responses, were lost interest in the intervention or research, the feeling that participating took too much time and the feeling that the intervention or research was not helpful.

Table 1. Baseline characteristics of participants ($N = 355$)

Characteristics	Featback ($N = 88$)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)
Gender					
Female (%)	82 (93.2)	89 (98.9)	84 (96.6)	88 (97.8)	343 (96.7)
Male (%)	5 (5.7)	1 (1.1)	1 (1.1)	2 (2.2)	9 (2.5)
Other (%)	1 (1.1)	0 (0.0)	2 (2.3)	0 (0.0)	3 (0.8)
Nationality					
Dutch (%)	78 (88.6)	80 (88.9)	80 (92.0)	81 (90.0)	319 (89.9)
Belgian (%)	9 (10.2)	9 (10.0)	6 (6.9)	8 (8.9)	32 (9.0)
Other (%)	1 (1.1)	1 (1.1)	1 (1.1)	1 (1.1)	4 (1.1)
Education					
Low (%)	5 (5.6)	12 (13.3)	12 (13.7)	18 (20.5)	47 (13.3)
Middle (%)	33 (37.5)	31 (34.4)	34 (39.0)	35 (39.3)	133 (37.6)
High (%)	50 (56.8)	47 (52.2)	41 (47.1)	36 (40.4)	174 (49.2)
Marital status					
Married/living together (%)	20 (22.7)	22 (24.4)	26 (29.9)	30 (33.3)	98 (27.6)
Living alone (%)	68 (77.3)	66 (73.3)	58 (66.7)	58 (64.4)	250 (70.4)
Divorced (%)	0 (0.0)	1 (1.1)	3 (3.4)	2 (2.2)	6 (1.6)
Widow (%)	0 (0.0)	1 (1.1)	0 (0.0)	0 (0.0)	1 (0.2)
Treatment history for ED					
Yes (%)	46 (52.3)	54 (54.0)	53 (60.9)	49 (54.4)	202 (56.9)
No (%)	42 (47.7)	36 (40.0)	34 (39.1)	41 (45.6)	153 (43.1)
Self-reported diagnosis status					
Officially diagnosed with ED	52 (59.1)	60 (66.7)	52 (59.8)	58 (64.4)	222 (62.5)
No diagnosis, but assumed to have ED	24 (27.3)	22 (24.4)	23 (26.4)	22 (24.4)	91 (25.6)
Eating problems, but likely no ED diagnosis	12 (13.6)	8 (8.9)	12 (13.7)	10 (11.1)	42 (11.8)
Age [Years]	28.0 (1.7)	28.3 (10.4)	26.8 (9.4)	28.1 (12.4)	27.8 (10.8)
Weight [kg]	64.0 (21.0)	62.2 (18.3)	63.6 (22.0)	64.7 (23.4)	63.6 (21.2)
Height [cm]	169.9 (7.2)	168.5 (6.9)	169.7 (7.1)	169.5 (6.9)	169.4 (7.0)
Duration of eating problems [years]	10.1 (9.1)	10.3 (8.8)	8.6 (8.2)	11.4 (12.0)	10.1 (9.7)

EDE-Q	3.9 (1.1)	4.1 (1.1)	4.3 (1.0)	4.3 (1.0)	4.1 (1.0)
PHQ-4	7.6 (3.4)	7.5 (3.3)	8.2 (2.9)	7.9 (3.3)	7.8 (3.2)
GSES	25.9 (5.8)	27.4 (5.2)	24.4 (5.4)	26.7 (5.8)	26.1 (5.6)
SSL-12	29.4 (6.7)	30.4 (7.5)	30.0 (6.7)	30.1 (7.0)	30.0 (7.0)
RSES	20.8 (5.5)	21.6 (5.8)	19.0 (4.5)	20.6 (4.9)	20.5 (5.3)
Motivation to change	21.4 (4.5)	22.6 (4.5)	22.2 (4.5)	22.0 (4.2)	22.0 (4.5)
Internet usage [hours per day]	4.2 (2.6)	3.7 (2.2)	3.9 (2.3)	3.4 (2.8)	3.8 (2.5)

Note. Data are presented as means (SD) unless indicated otherwise.
ED = Eating Disorder; EDE-Q = Eating Disorder Examination Questionnaire; GSES = General Self-Efficacy Scale; PHQ-4 = 4-item Patient Health Questionnaire; RSES = Rosenberg Self-Esteem Scale; SD = standard deviation; SSL-12 = 12-item Social Support List.

Eating disorder psychopathology

Main results can be found in Table 2 and all mixed model results can be found in Appendix B.2. First, the three active interventions were compared with the waiting list control condition (CC1*TC1). A statistically significant medium effect of condition on changes in EDE-Q total scores between baseline and post intervention was found, favoring the active interventions. There were no significant condition-by-time interaction terms for the other time contrasts, indicating that there were no differences in longer-term eating disorder symptom changes between the three active interventions and the waitlist. No other significant results were found regarding eating disorder symptomatology. Specifically, the combination condition did not outperform the Featback only and expert-patient support only conditions pooled together in the short term (CC2*TC1) or any other time contrast. Similarly, no difference in effectiveness between the Featback only and expert-patient support only conditions were found in the short term (CC3*TC1) or any other time contrast.

To further explore change in participants on EDE-Q scores between baseline and post intervention, the reliable change index (RCI) was calculated (Jacobson & Truax, 1992). Based on the EDE-Q reliability in the current sample (Cronbach's $\alpha = .90$) and the standard deviation of baseline EDE-Q total scores (1.04), the RCI was 0.89. Derived from the RCI, the number of participants (averaged across 100 imputed datasets) showing reliable deterioration, no change and reliable improvement was 14, 261 and 80 respectively. No significant difference in these frequencies between conditions was found, $\chi^2(6) = 11.14$, $p = .08$. Details can be found in Appendix B.3.

Secondary outcomes

For symptoms of anxiety and depression, social support and self-efficacy, no time by condition interaction effects were found, indicating that for these variables trajectories over time were similar across conditions. Between baseline and post intervention, participants improved regarding symptoms of anxiety and depression. Between post intervention and 12-month follow-up participants improved regarding self-efficacy. Results were non-significant across other condition and time contrasts.

Figure 1. Participant flow during the study.

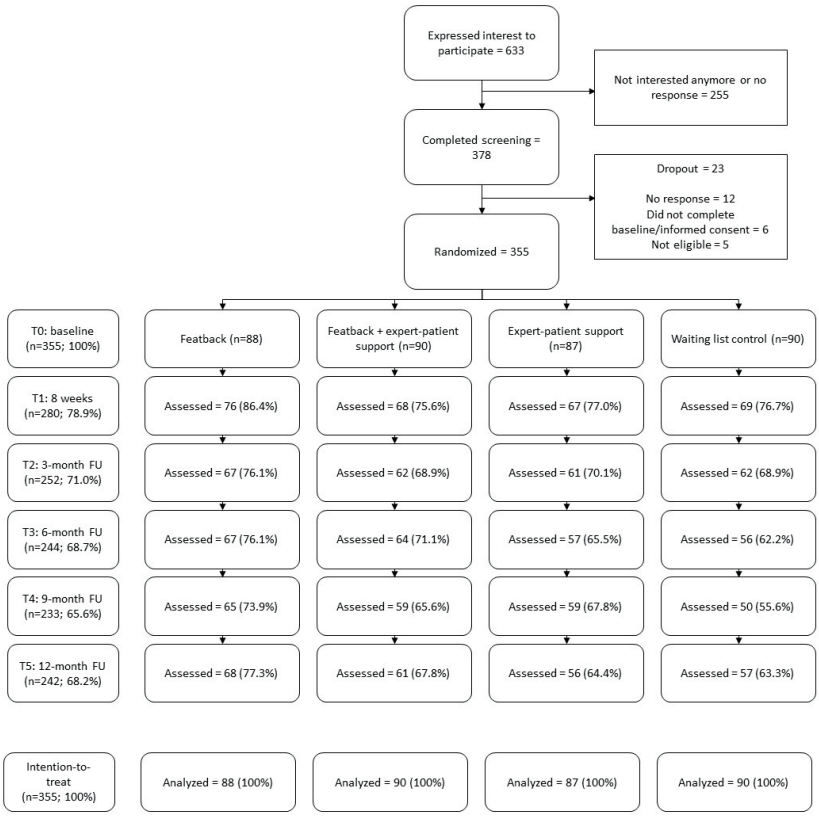
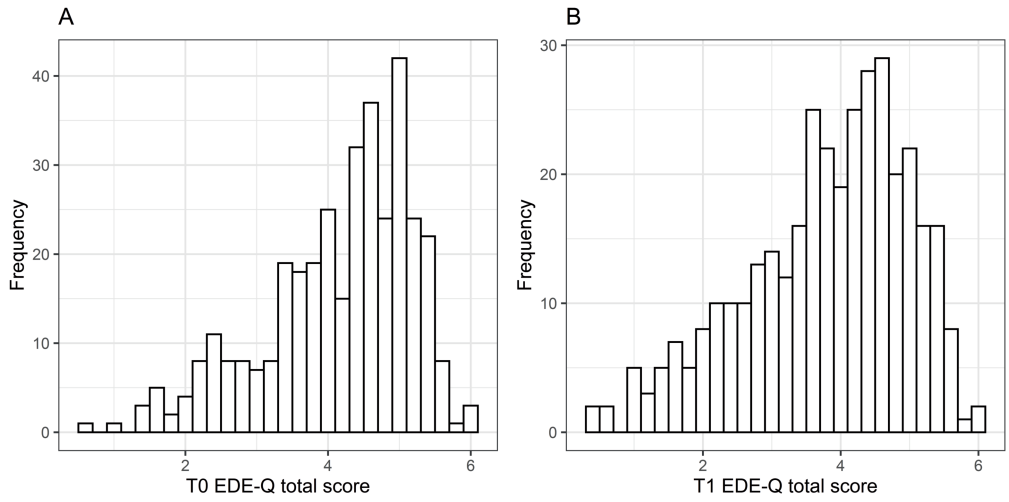


Table 2. Pooled results of linear mixed models analyses over 100 multiple imputed datasets

Measure	CC	TC	Time effects		Time-condition interaction effects		
			β (95% CI)	t (p)	β (95% CI)	t (p)	Cohen's d
Eating disorder symptoms (EDE-Q global)	CC1	TC1	-0.18 (-0.22; -0.14)	-8.12 ($< .001$)	-0.15 (-0.22; -0.07)	-3.66 ($< .001$)	0.38
	CC1	TC5	-0.27 (-0.35; -0.19)	-6.67 ($< .001$)	0.16 (0.02; 0.29)	2.26 (.02)	0.25
	CC2	TC1	-0.23 (-0.28; -0.18)	-8.81 ($< .001$)	-0.04 (-0.12; 0.03)	-1.17 (.24)	-0.12
	CC2	TC5	-0.22 (-0.31; -0.13)	-4.87 ($< .001$)	-0.04 (-0.18; 0.09)	-0.64 (.52)	-0.09
	CC3	TC1	-0.21 (-0.26; -0.15)	-6.73 ($< .001$)	0.01 (-0.06; 0.07)	0.17 (.87)	0.02
	CC3	TC5	-0.2 (-0.31; -0.09)	-3.66 ($< .001$)	0.01 (-0.09; 0.12)	0.25 (.80)	0.04
Depression and anxiety symptoms (PHQ-4)	CC1	TC1	-0.41 (-0.58; -0.23)	-4.61 ($< .001$)	-0.22 (-0.52; 0.08)	-1.43 (.15)	-0.12
	CC1	TC5	-0.42 (-0.63; -0.20)	-3.85 ($< .001$)	0.16 (0.02; 0.29)	2.26 (.02)	0.25
General self-efficacy (GSES)	CC1	TC1	0.09 (-0.18; 0.35)	0.65 (.52)	0.09 (-0.4; 0.59)	0.37 (.71)	0.04
	CC1	TC5	0.38 (0.07; 0.69)	2.40 (.02)	-0.15 (-0.73; 0.43)	-0.50 (.62)	-0.06
Experienced social support (SSL-12)	CC1	TC1	-0.04 (-0.35; 0.26)	-0.28 (.78)	-0.19 (-0.72; 0.35)	-0.69 (.49)	-0.07
	CC1	TC5	0.35 (-0.1; 0.80)	1.52 (.13)	0.20 (-0.58; 0.97)	0.50 (.62)	0.06

CC = Condition contrast, CI = confidence interval, TC = Time contrast
CC1 = Three active interventions (Featback only, expert-patient support only and Featback plus expert-patient support) versus waiting list control condition
CC2 = Featback plus expert-patient support condition versus Featback only and expert-patient support only
CC3 = Featback only versus expert-patient support only
TC1 = baseline versus post intervention
TC5 = post intervention versus 12-month follow-up

Figure 2. Distribution of eating disorder examination questionnaire (EDE-Q) total scores at baseline (A) and post intervention (B)



Intervention check

Raters could distinguish sessions correctly in 94% of the cases (agreement between the two raters was 95%), confirming the expectation that differences exist between the psychologist sessions of the previous trial (Aardoom, Dingemans, Spinhoven, et al., 2016) and expert-patient sessions of the current trial. Details on the intervention check can be found in Appendix B.4.

Similarity ratings

Of the 177 participants who had the option to receive expert-patient support, 144 (81.4%) completed the IOS-scale. Answers on the IOS-scale were not imputed, since only participants in the two expert-patient support conditions received this questionnaire. Perceived similarity ratings were low with a median of 2 and mean of 2.7 ($SD = 1.5$). Perceived similarity was not predictive of eating disorder symptoms at post intervention, $\beta = 0.002$, $F(1, 126) < 0.001$, $p = .98$, and, after correcting for multiple testing, at 12-month follow-up, $\beta = -0.19$, $F(1, 107) = 4.05$, $p = .05$. Similar non-significant results were found for the predictive value of perceived similarity ratings on experienced social support and self-efficacy.

Per protocol analyses

The number of adequate-dose participants was 74 (84.1%) in the Featback, 34 (37.8%) in the combination and 48 (55.2%) in the expert-patient support condition. The main analysis,

concerning the comparison of the effect of the three active interventions and the waitlist condition on eating disorder symptoms for the period between baseline and post intervention, was repeated with adequate-dose participants ($N = 156$). The result was similar to the main analysis and favoring the three active interventions, $t(323) = -3.23$, $p = .001$. The number of intervention sessions was not prognostic of eating disorder symptoms at post intervention, $\beta = -0.02$, $t(665) = -1.10$, $p = .27$. Furthermore, adding covariates to the model yielded similar results to the main analysis, with a significant time-by-condition interaction favoring the three active interventions, $t(405) = -3.57$, $p < .001$. Lastly, severity of eating disorder symptoms was explored as a moderator, by entering it as a fixed effect in the CC1*TC1 model. Based on (Mond et al., 2006), severity was considered high or low depending on whether a participant's baseline EDE-Q score was higher or lower than 4.0. No evidence for moderation of eating disorder symptom severity was found, $t(447) = 0.80$, $p = .42$.

Table 3 presents information on intervention usage, satisfaction and initiation of professional help. Further details can be found in Appendix B.5. On average, participants made more use of Featback sessions than expert-patient support sessions. Intervention satisfaction was significantly higher in conditions with expert-patient support compared to the Featback only condition. Furthermore, of the 150 participants seeking help at post intervention, 33 (22%) indicated that the 8-week intervention stimulated them to request professional help. However, there was no difference between the four conditions in initiating professional help for eating-related problems (e.g., with general practitioner or psychologist). Lastly, of the 90 participants randomized to the waiting list, 38 (42%) made use of the option to use Featback and expert-patient support after study completion.

Table 3. Intervention usage, satisfaction and help initiation of participants per condition

Category	Featback ($N = 88$)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waitlist ($N = 90$)	Total sample ($N = 265$; excludes waitlist)	Statistics
Adequate-dose participants (%)	74 (84.1)	34 (37.8)	48 (55.2)	NA	156 (58.9)	$\chi^2(2) = 40.00$, $p < .001$
Indicated to stop with the intervention (%)	4 (4.6)	3 (3.3)	6 (6.9)	NA	13 (4.9)	$\chi^2(2) = 1.24$, $p = .54$
Mean number of sessions (SD; median)	6.5 (2.1; 7.0)	9.2 (5.2; 10.0)	4.4 (3.1; 5.0)	NA	6.7 (4.2)	$F(2, 262) = 37.67$, $p < .001$
Featback		5.6 (2.7; 7.0)		NA	6.0 (3.0)	$t(176) = 2.28$, $p = .024$
Featback		3.6 (2.9; 3.0)		NA	4.0 (3.0)	$t(175) = 1.86$, $p = .066$
Proportion of possible sessions used (SD)	80.7% (26.1)	57.4% (32.2)	54.9% (38.3)	NA	64.3% (34.5)	$F(2, 262) = 16.80$, $p < .001$
0 sessions (%)	2 (2.3)	6 (6.7)	17 (19.5)	NA		$\chi^2(2) = 16.43$, $p < .001$
Featback		6 (6.7)		NA		
Support		23 (25.6)		NA		
1 session (%)	4 (4.5)	4 (4.4)	7 (8.0)	NA		$\chi^2(2) = 1.38$, $p = .50$
Featback		6 (6.7)		NA		
Support		5 (5.6)		NA		
All sessions (%)	40 (45.5)	11 (12.2)	14 (16.1)	NA		$\chi^2(2) = 16.43$, $p < .001$
Featback		34 (37.8)		NA		
Support		12 (13.3)		NA		
Intervention satisfaction[a] (SD)	5.8 (1.8)	7.1 (1.7)	7.4 (1.6)	NA	6.7 (1.8)	$F(2, 192) = 15.98$, $p < .001$
Initiated professional help[b] (%)	22 (25.0)	17 (18.9)	15 (17.2)	20 (22.2)	74 (20.8)	$F(3, 351) = 0.64$, $p = .59$

Exp = expert-patient support; FB = Featback SD = standard deviation; WLC = waiting list control condition
[a] On a scale from 1 (completely unsatisfied) to 10 (completely satisfied); based on 70, 63, and 62 (195 total) responses in the Featback, Featback + expert-patient support and expert-patient support conditions respectively.
[b] Indicates the number of participants who had never received eating disorder related treatment at baseline, but sought professional help (e.g., with a general practitioner, psychologist or psychiatrist) on at least one of the follow-up measures (T1-T5); total based on 355 participants.

Discussion

The main aim of the current study was to investigate the effectiveness of Featback, expert-patient support and the combination of both compared to care as usual (waitlist) for eating

disorders. Of the 355 participants, 22.5% (25.3% in the three active conditions and 13.3% in the waitlist) experienced reliable improvement in eating disorder symptoms. Results showed that participants in the three active conditions had larger improvements in eating disorder symptoms over the 8-week intervention period than participants in the waitlist condition. Contrary to expectations, the three active interventions were equally effective. The conditions with support from an expert patient were rated as more satisfactory than Featback alone. The results are similar to a previous randomized controlled trial on the effectiveness of Featback (Aardoom, Dingemans, Spinhoven, et al., 2016), where active interventions outperformed a waiting list control condition and adding support from a psychologist to the automated monitoring system enhanced intervention satisfaction, but not its effectiveness. Results are also in line with other research that found internet-based interventions to be effective in reducing eating disorder symptoms (Linardon et al., 2020; Melioli et al., 2016). The interventions in the current trial did not lead to more professional treatment initiation compared to the waitlist condition, but over 30 participants indicated the intervention had stimulated them to seek professional help. That a brief intervention such as Featback and expert-patient support can improve eating disorder symptoms and might stimulate help-seeking behaviors is promising, especially when considering its potential reach. Featback is free to use and easily accessible 24 hours a day. Therefore, it can exist next to and complement current treatment options for eating disorders, by reaching underserved individuals (Bauer & Goldschmidt, 2019). This is supported by the fact that 43% of the current study sample had, at baseline, never received treatment while having a very long average duration of eating disorder problems. This suggests that others in similar positions might also profit from Featback or expert-patient support, even if they do not (yet) receive other forms of treatment. Apart from individuals not currently reached by regular treatment, Featback could also be used in the period between intake and commencement of treatment. Waiting times for eating disorder treatment in the Netherlands have increased to an average of 11 weeks (Nederlandse Zorgautoriteit, 2021). Although further research is necessary, offering self-help interventions in this period might keep individuals motivated for recovery and prepare them for future care (Vollert et al., 2019). In general, based on findings from this and other studies, low-threshold and innovative interventions like Featback and expert-patient support are not only likely to reduce the burden for individuals with eating disorders, but also to bridge the treatment gap by reaching underserved people.

Expert-patient guidance

It was expected that individuals with an eating disorder would bond quickly with expert patients, making them receptive to interventions aimed at changing destructive behavior and increasing self-efficacy and experienced social support (Dennis, 2003; Simoni et al., 2011). Contradictorily, adding expert-patient support to Featback did not increase effectiveness. This is not in line with a meta-analysis, covering multiple disorders, suggesting a small beneficial effect of guidance (Baumeister et al., 2014). Guidance in the pooled studies varied considerably and mostly covered support by therapists. An explanation of the discrepancy might thus be that expert patients are less effective in reducing symptoms through online support than health professionals. However, even though expert-patient and psychologist support appear to be distinct interventions, a previous Featback trial found no increased ef-

fectiveness of adding online psychologist support (Aardoom, Dingemans, Spinhoven, et al., 2016). Alternatively, results on the effectiveness of online guidance of Baumeister et al. (2014) may not generalize to eating disorder interventions, where evidence seems mixed (Yim & Schmidt, 2019b). Accordingly, in a meta-analysis studying the effect of individual components on effectiveness of e-health interventions for eating disorders, guidance did not moderate intervention effectiveness (Barakat et al., 2019). Regardless of effectiveness, individuals with eating disorders are repeatedly found to value support in the context of internet interventions highly (Linardon et al., 2021; Yim & Schmidt, 2019a). It suggests that intervention effectiveness does not require satisfaction. Surprisingly, despite the higher satisfaction, there were relatively fewer adequate-dose participants in conditions with expert-patient support. For the combination condition, this may partly be explained by the fact that an adequate dose required 10 completed sessions instead of 5 in the other conditions. Nonetheless, many participants in the current study did not make (full) use of the option to receive weekly support, indicating they still experienced barriers to engage with the online support. These barriers were thought to be lower when guidance was offered by expert patients compared to health professionals. Specifically, perceived similarity between the participant and expert patient was proposed to be an important ingredient for the effectiveness of expert-patient support, but no proof for this was found. Possibly, the perceived similarity was too low and little varied to detect any effect or three weeks were too few to build a rapport.

Future directions

Considering the ambiguity around guidance for internet-based eating disorder interventions, more research specifically devoted to how it works and under which circumstances is warranted. Additionally, investigating predictors and moderators of internet-based interventions like Featback and expert-patient support might clarify who benefits from what kind of intervention, leading to more personalized treatment. Lastly, an interesting next step might be to see if Featback and expert-patient support can be improved by incorporating more innovative technologies, as there is still much ground to cover (Burger et al., 2020). For example, gamification, videos or Virtual Agents in addition to text alone show potential in improving mental health and engagement (Abd-Alrazaq et al., 2020; Fleming et al., 2017). This is in line with evidence indicating that using multiple features that address different modalities has a positive influence on the effectiveness of technology-enhanced eating disorder treatments (Barakat et al., 2019). Lastly, to better understand the effect of intervention usage on effectiveness, it might be valuable to investigate different durations of Featback and expert-patient support.

Strengths and limitations

Strengths of the study are recruitment of a large sample size, the design including randomized allocation of participants, obtaining 12-month follow-up measures in all conditions with adequate retention of participants, and maintaining an intent-to-treat approach making use of multiple imputations of the data. A limitation is the sole use of self-report measures, which are subject to socially acceptable answers, misinterpretation, and recall bias. Nevertheless,

using self-report allowed for maintaining the low-threshold character of the intervention and participating in the study. Similarly, broad inclusion criteria were used to ensure the sample represented intended end users, improving generalization to real-world settings. However, it is difficult to generalize the results to individuals with specific diagnoses, as rates of improvement may differ across diagnostic groups. A last consideration pertains to the number of planned contrasts. While they were conform the hypotheses as stated in the protocol, performing several tests increased the family-wise error rate, which was mitigated by a Bonferroni correction of the p-values. Nevertheless, a different statistical approach (e.g., factorial design with main and interaction effects of treatment) might have involved fitting fewer models, while still informing on the (relative) effectiveness of the internet interventions.

Conclusion

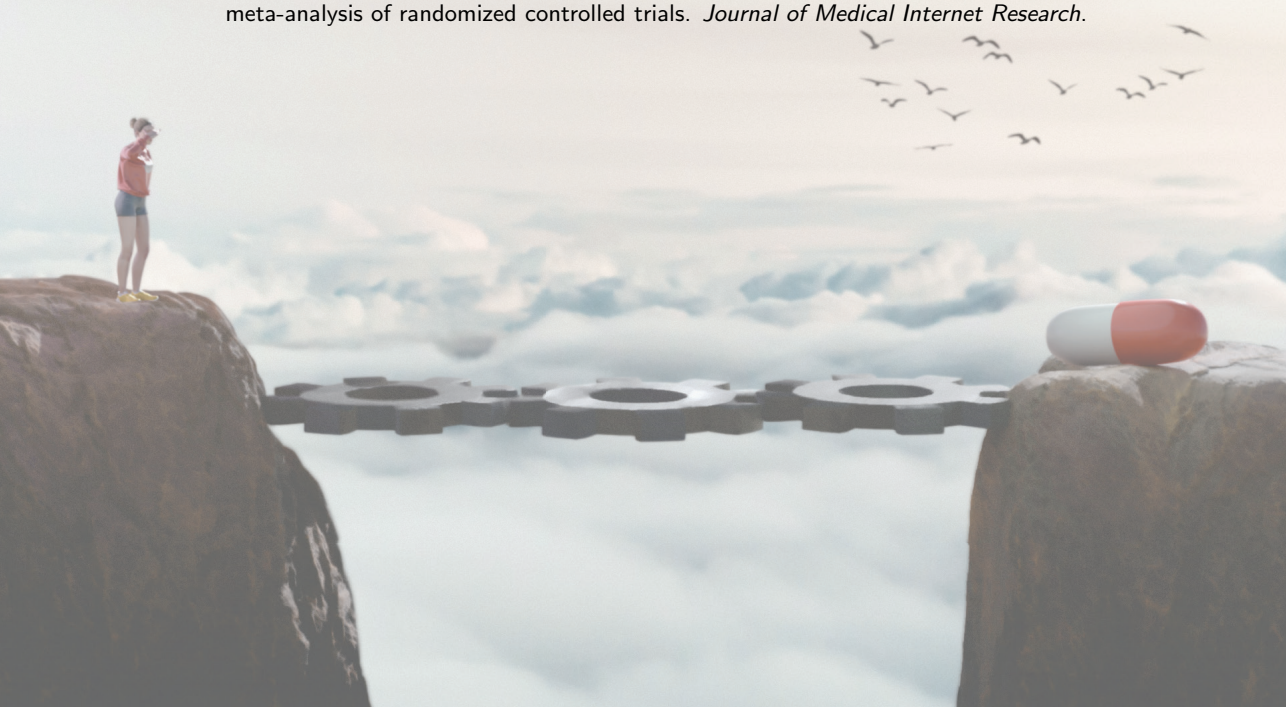
A fully automated low-threshold internet-based self-help program for eating disorders (Featback), weekly chat or email support from an expert patient and the combination of both were effective in reducing eating disorder symptoms compared to a waiting list control condition. Although expert-patient support improved satisfaction ratings, it did not improve the effectiveness. Now that beneficial effects of Featback have been confirmed in two randomized controlled trials, a next step is to implement the program and make it widely available. The current study highlights the potential of internet interventions such as Featback and expert-patient support to reach the large group of undetected and underserved individuals with an eating disorder and help them address their problems, complementing the existing pallet of treatment options that currently exists for eating disorders.

Chapter 4

Cost-effectiveness of internet interventions compared to treatment as usual for people with mental disorders: a systematic review and meta-analysis of randomized controlled trials

Trial Registration: Prospero registration https://www.crd.york.ac.uk/PROSPERO/; registration number: CRD42019141659.

Accepted for publication: Rohrbach, P. J., Dingemans, A. E., Evers, C. A. J. M., Aardoom, J. J., Van Furth, E. F., Spinhoven, P., Lähde, I., Clemens, F., & Van den Akker-Van Marle, M. E. (in press). Cost-effectiveness of eHealth interventions compared to treatment as usual for people with mental disorders: a systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*.



Abstract

Background: Economic costs of mental disorders for society are huge. Internet-based interventions are often coined as cost-effective alternative to usual care, but evidence is mixed.
Objective: The aim was to review the literature on cost-effectiveness of internet interventions for mental disorders compared to usual care and to provide an estimate of the monetary benefits of such interventions compared to usual care. **Methods:** A systematic review and meta-analysis of (1) randomized controlled trials that (2) included participants with symptoms of mental disorders, (3) investigated a telephone or internet-based intervention, (4) included a control condition in the form of treatment as usual, psychological placebo, waiting list control or bibliotherapy, (5) reported outcomes on both quality of life and costs and (6) were published in English, was conducted. Electronic databases PubMed (including Medline), Embase, Emcare, PsycINFO, Web of Science and The Cochrane Library were used. Data on risk of bias, quality of the economic evaluation, quality adjusted life years (QALYs) and costs were extracted from included studies and the incremental net benefit (INB) was calculated and pooled.
Results: The search yielded 6226 abstracts and 37 studies with 14,946 participants were included. Quality of economic evaluations of included studies was rated to be moderate and risk of bias was high. A random-effects approach was maintained. Analyses suggested internet interventions to be slightly more effective than usual care in terms of QALY gain, Hedges' $g = 0.05$ (95% CI 0.01; 0.10, $P = .016$), and equally expensive, Hedges' $g < 0.01$ (95% CI -0.08; 0.84, $P = .96$). The pooled INB was \$255, (95% CI \$91; \$419, $P = .002$), favoring eHealth interventions over usual care. Perspective of the economic evaluation and targeted mental disorder moderated results.
Conclusions: Findings indicate that cost-effectiveness of internet interventions for mental disorders compared to a care-as-usual approach is likely, but generalizability to new studies is poor given the substantial heterogeneity. This is the first study in the area of mental health to pool cost-effectiveness outcomes in an aggregate-data meta-analysis.

Introduction

Mental disorders have a huge impact on sufferers as well as on society. It is appraised to cause almost one-third of global years lived with disability and account for roughly 10% of all disability-adjusted life years, placing it in the top 3 causes of global burden worldwide (Vigo et al., 2016). When mental, neurological and substance use disorders are taken together, the global economic costs in 2010 were estimated to be 2.5 trillion US dollars, with projections for 2030 being around 6 trillion dollars (Marquez & Saxena, 2016). While these numbers are serious and account for around 2.3-4.4% of the gross domestic product in high-income countries, most countries spend a disproportionally small amount of their health budget on mental health (Organisation for Economic Co-operation and Development, 2014). This warrants changes in policy, but also stresses the need for effective and inexpensive interventions so that individuals can recover more swiftly from a mental disorder and the global burden is reduced.

Effectiveness of internet interventions

Swift technological advancement brings the promise of new and effective interventions. Indeed, internet interventions have become a popular niche of research and treatment. A reason for attempting to create effective internet interventions is to reduce the treatment gap, which specifies the discrepancy between the proportion of people who need help for a particular disorder and the proportion of those individuals who actually receive care (Bennett & Glasgow, 2009; Kazdin et al., 2017). In other words, internet interventions can be used to reach an underserved population of people with a (risk for developing) mental illness (Aardoom, Dingemans, & Van Furth, 2016). Internet interventions for people with (symptoms of) a mental disorder are increasingly confirmed in their effectiveness. There have been several meta-analyses on this topic covering various mental disorders, such as depression, (social) anxiety, post-traumatic stress and eating disorders (Andersson et al., 2014; Carlbring et al., 2017; Melioli et al., 2016; Sander et al., 2016; Välimäki et al., 2017). A recent umbrella review of meta-analyses shows that there is sufficient information now to assume the effectiveness of internet interventions (Andersson et al., 2019). In general, guided internet interventions (most of which have a cognitive-behavioral underpinning) seem to be as effective in reducing symptomatology as face-to-face treatment and outperform waiting list control conditions. Unguided internet interventions seem to be more effective than waiting list control conditions, but less effective than guided internet or face-to-face interventions.

Cost-effectiveness of internet interventions

Internet interventions for mental disorders are often coined as a cost-effective alternative to established treatments (F. Griffiths et al., 2006), but results on cost-effectiveness are tentative at best (Hedman et al., 2012; Tate et al., 2009). Individual studies show mixed results and their heterogeneity in methods, outcomes and comparators makes it difficult to draw conclusions (Donker et al., 2015; Naslund et al., 2022), so no definitive assumptions can yet be established. Additionally, cost-effectiveness studies that are conducted alongside randomized controlled trials (RCTs) are often not powered for economic evaluations, limiting

their predictive ability (Hollingworth et al., 2013). Nevertheless, the separate studies make an important contribution to the rapidly growing body of evidence, so that it becomes more feasible to make meaningful overviews in the form of systematic reviews and meta-analyses. Indeed, Naslund et al. (2022) performed a broad systematic review on cost outcomes for telemedicine interventions for mental disorders and Donker et al. (2015) systematically reviewed RCTs on cost-effectiveness of internet-based interventions for mental disorders. Both reviews concluded that internet interventions for mental disorders have the potential to be cost-effective compared to alternatives, but evidence is still circumstantial. Kolovos et al. (2018) performed an individual participant meta-analysis to investigate the cost-effectiveness of internet interventions compared to a control condition (e.g., waiting list or care as usual) for depression. The authors cautiously conclude that results showed no indication of cost-effectiveness and remark that adding economic evaluations to trials more frequently would help to reach well-founded deductions.

Pooling cost-effectiveness data

To get a precise estimate of the cost-effectiveness of internet interventions for mental disorders compared to alternatives it is desirable to pool outcomes of individual studies in an aggregate data meta-analysis. As cost-effectiveness is expressed as a combination of two variables, the difference in costs and effects between an intervention and control condition, this is statistically complex. Fortunately, Crespo et al. (2014) developed a theoretical framework to do these kinds of analyses. This method has been successfully applied by another research team in several studies in different areas of medicine (Bagepally et al., 2020; Bagepally et al., 2019; Chaiyakittisopon et al., 2021; Haider et al., 2019; Noparatayaporn et al., 2021). Currently, no meta-analysis on the cost-effectiveness of internet interventions for mental disorders has been conducted. The individual participant data (IPD) meta-analysis by Kolovos et al. (2018) looked at depression only. The IPD approach is more reliable (Riley et al., 2010), but decreasingly feasible when more studies are included. Furthermore, the last overview of RCTs on cost-effectiveness of internet-based interventions for mental disorders was in 2015 (Donker et al., 2015), but results were not pooled in a meta-analysis.

Concordantly, given the mixed evidence and limitations of individual studies and the rapid increase of novel research, a thorough and recent overview of the literature on the cost-effectiveness of internet interventions for individuals with (symptoms of) a mental disorder compared to alternatives is warranted. A first step would be to establish whether internet interventions are cost-effective compared to control conditions such as a waiting list and care as usual. The current article, then, aims to investigate the cost-effectiveness of internet interventions for mental disorders compared to control conditions by conducting a systematic literature search and aggregate data meta-analysis of RCTs.

Methods

The literature review was conducted according to the Preferred Reporting Items for Systematic Review (PRISMA) guidelines (Moher et al., 2015), see Appendix C.1. This review was registered in Prospero (registration number CRD42019141659).

Eligibility criteria

Studies were included that (1) were RCTs; (2) included participants (of all ages) who have symptoms of or a diagnosed mental disorder; (3) investigated a telephone- or an internet-based (that work via computer, tablet and/or smartphone) intervention. All forms of internet-based interventions were considered, including fully automated (i.e., unguided) interventions, guided interventions and teleconferencing interventions. Guided interventions could contain asynchronous support (i.e., a delay in the support such as with mail or forum services) or synchronous support (i.e., no delay in the support such as with videoconferencing and chat). Additionally, smartphone apps with the purpose of elevating symptoms of a mental disorder were included; (4) included a control condition in the form of (enhanced) treatment as usual, psychological placebo, waiting list control or bibliotherapy; (5) reported outcomes on both quality of life (quality adjusted life years; QALYs) and costs; (6) were published in English. It is worth noting that studies were included if participants had somatic conditions (e.g., cancer or diabetes) as long as the participants had comorbid symptoms of a mental disorder and the investigated intervention had as primary aim to alleviate these symptoms. Studies were excluded if the main intervention exclusively relied on wearable devices or virtual reality, had only a face-to-face intervention as control condition, or did not provide sufficient information on costs or QALYs.

Search strategy and selection criteria

Electronic databases PubMed (including Medline), Embase, Emcare, PsycINFO, Web of Science and The Cochrane Library were searched up until 1 March 2021. A search string was made for PubMed and translated for the other databases. The PubMed search string contained MeSH terms for the concepts of 'mental disorders', 'cost-benefit analysis' and 'telemedicine'/'internet', with all expressions under these headings included as free terms as well. Furthermore, other terms related to the three MeSH terms were added to maximize the sensitivity of the search. The full PubMed search strategy can be found in the Appendix C.2. By checking cross-references in the included studies we minimized the chance of missing relevant data. During the screening phase, all relevant study protocols, conference abstracts and trial registrations were identified and authors of unpublished studies with potentially relevant data were contacted.

The identified articles were all screened in three steps and by two researchers (PR, AD, CE, EA, IL or FC). First, the title and abstract of the eligible articles were screened. Subsequently, the full texts of all the included abstracts were screened for eligibility. Lastly, the relevant data were extracted. If there was any disagreement between the two researchers in any of the three steps a third researcher of the team made the final decision.

Quality assessment

Risk of bias of the included RCTs was assessed by two researchers (PR, AD, CE, IL or FC) using the Cochrane Risk of Bias assessment tool (Higgins et al., 2011). Specifically, data were gathered on the topics of (1) random sequence generation (selection bias), (2) allocation concealment (selection bias), (3) blinding of participants and personnel (performance bias),

(4) blinding of outcome assessment (detection bias), (5) incomplete outcome data (attrition bias), (6) selective reporting (reporting bias), (7) other sources of bias. Finally, for all these topics the risk of bias was assessed (i.e., low, unclear or high risk of bias). Any disagreement between the researchers on (each of the seven areas of) the risk of bias was resolved by means of a discussion between the two raters or by a third rater. A final rating of high, medium or low bias was assigned to each study based on the revised tool for assessing risk of bias (Sterne et al., 2019). Specifically, high risk of bias was assigned to studies when a high risk of bias in any domain was present or unclear risk of bias was present in two or more domains; medium risk of bias was assigned in the case of one unclear rating across all domains; low risk of bias was assigned when all domains were rated as low risk of bias. Exceptions were that blinding of participants and personnel was not considered in the final risk of bias rating as blinding was unfeasible in most studies, and unclear risk of bias on the selective reporting domain was considered as low risk of bias for the final risk of bias rating as the absence of a published protocol or trial preregistration is still common.

Furthermore, the 19-item CHEC list (Evers et al., 2005) was used to assess the quality of the economic evaluation for all included studies. The expert on economic evaluations (EA) scored all articles on the CHEC list, while other authors (PR, AD, CE, IL and FC) divided included articles for the CHEC list as well, so articles were rated twice. All discrepancies between raters on the CHEC list were resolved by means of a discussion. Assigning an overall quality score for the CHEC list is not advocated, as cutoff scores are highly heterogeneous (Watts & Li, 2019) and difficult to substantiate. Since we wanted to group studies by quality of the economic evaluation in analyses, we adopted an approach using a selection of items of the CHEC list. A study had to fulfill at least 8 specific items (i.e., items 7-14, see Table 4 for a list of all items) to be deemed of high quality. These items were chosen, because they contribute to the assessment of cost and effects used in the current meta-analysis. Other items (i.e., 1-4 and 15-19) were deemed less important for this study, as they aim at clarifying the text and the appropriateness of the used methods and analyses, which is also captured partly in the risk of bias assessment. Finally, items 5 and 6 regarding time horizon and perspective were not considered for the final quality rating, because these were already explored in moderator analyses.

Outcome measures

Quality of life

The difference in the average number of QALYs gained per participant in the intervention group and its comparator (delta QALY) was the target health outcome measure. A QALY indicates one extra life year in perfect health. QALYs are derived from generic health-related quality-of-life measures (e.g. EQ-5D or 36-item Short Form Health Survey; SF-36) that are transformed into utility scores, multiplied by the time (in years) a participant spent with that utility (Singh et al., 2001).

Costs

Delta societal costs or delta healthcare costs (from now on referred to as delta costs), which indicate the difference in costs between the intervention and control condition, was the

primary outcome measure. Studies with a healthcare perspective estimated costs per study group by measuring healthcare use of participants during the intended follow-up period and multiplying that with a reference price for the used services. Studies with a societal perspective also included costs based on, for example, absence from paid and unpaid work, reduced productivity while at work (presenteeism) and/or domestic care for participants.

Two steps were taken to account for cost differences. First, inflation was controlled for by recalculating all costs to the level of 2021 using consumer price indexes as reported for the countries in which each study was conducted (OECD, 2021b). Second, similar articles or services have different prices between countries, indicating differences in purchasing power. Purchasing Power Parities (PPP) were used to transform costs according to the most recent rates from 2017 (OECD, 2021a), effectively converting all costs to US dollars. A general indexing factor was used for PPP rather than a healthcare specific one, since other costs were also involved when studies employed a societal perspective.

Data preparation

All data extraction items can be found in the Appendix C.3. Apart from the outcome measures needed for the meta-analysis, other data were extracted either for sensitivity analyses or for further exploration. Data needed for the meta-analysis were (1) delta QALY, (2) variance of delta QALY, (3) delta costs, (4) variance of delta costs, and (5) covariance of delta QALY and delta costs. With these variables available for each study it was possible to calculate the incremental net benefit (INB) and pool all INBs using the method as described by Crespo et al. (2014). The INB indicates gains of an intervention compared to another expressed in monetary terms. Delta QALY and delta costs were either retrieved directly from the articles or calculated by subtracting the average QALYs or costs per patient in the intervention condition from those in the control condition. The method to retrieve the variances and covariance was based on that of Bagepally et al. (2019) and described here in order of the most ideal (i.e., reliable) to least ideal scenario. Appendix C.4 lists all used formulas for this meta-analysis including the INB.

Calculating the variance of delta QALY

Scenario 1. Studies report the standard deviation (SD), standard error (SE) or 95% confidence interval (CI) of delta QALY. The variance can directly be calculated by formula 1 or 2. When the 95% CI is reported an additional step is required, for which we used formula 3.

Scenario 2. Studies report the SD, SE or 95% CI of QALYs for the separate conditions (but not for the difference). If the SDs of separate conditions are reported the variance of the difference between two conditions can be calculated using formula 4. If the SEs of separate conditions are reported the variance can be calculated using formula 5. If the 95% CIs of separate conditions are reported the SE is first calculated using formula 3.

Scenario 3. Studies report only the average of separate conditions, but no measure of spread. In this case, first the corresponding author of the article was contacted to inquire about the possibility of receiving an indication of spread (SD, SE or 95% CI) of delta QALY or of the spread of the average QALY gain per condition. If this was not possible or the authors did not respond, two further options remained to estimate the measure of spread.

First, if a cost-utility plane with bootstrapped delta QALY and delta healthcare or societal costs was reported, the free available software Webplot Digitizer (Rohatgi, 2020) was used to reverse engineer the individual data points. The points were exported to an Excel file, from which the SD of delta QALY could be calculated. An example of how the software was used can be found in the Appendix C.5. Webplot Digitizer has been used in various studies and is found to be a reliable tool for extracting data with high intercoder reliability and validity (Burda et al., 2017; Drevon et al., 2017). However, the precision of the software seems to be dependent on the visual presentation of individual graphs (Moeyaert et al., 2016). Second, if no cost-effectiveness plane was reported the measure of spread of delta QALY was estimated by taking the mean of the two most similar studies or comparisons in terms of delta QALY, number of participants and investigated intervention.

Calculating the variance of delta costs

For these calculations identical steps were followed as for the variance of delta QALY, but using costs instead of QALYs.

Calculating the covariance between delta QALY and delta costs

To estimate the covariance between delta QALY and delta costs a bootstrap procedure is necessary to be able to have multiple estimates of delta QALY and delta costs. The covariance was never reported in included articles. As shown by (Bagepally et al., 2019), an approximation without the original data is possible, albeit less reliable. Hence, all corresponding authors of included studies were contacted and asked to provide the covariance, to rely on author data for estimates of this parameter as much as possible.

Scenario 1. Authors were able to provide the covariance or information needed to calculate it (e.g., the data set including all bootstrapped delta QALYs and delta costs or individual participant data needed to perform a bootstrap procedure including (1) the allocated condition of participants, (2) the QALYs over the entire follow-up period and (3) the costs over the entire follow-up period).

Scenario 2. If authors did not respond or could not provide information on the covariance, but a cost-utility plane was presented in the article Webplot Digitizer was used. The covariance could be calculated after using the software to estimate the data points (delta QALYs and delta costs).

Scenario 3. If authors did not respond or could not provide information on the covariance and no cost-utility plane was presented in the article, a different approach was used. First, the mean correlation between delta QALY and delta costs was calculated for all studies where the covariance was obtained with data received directly from authors (i.e., covariances obtained from scenario 1) using formula 6. Second, the mean correlation was used to calculate the covariance between delta QALY and delta costs for studies falling under scenario 3 using formula 6 (mean imputation).

Statistical analyses

Data preparation was done in Excel and analyses were conducted in Comprehensive Meta-Analysis software, version 3 (Borenstein et al., 2013). First, we calculated the INB for each study. The first step is to multiply society's willingness to pay (WTP) for one extra year lived in perfect health (i.e., 1 QALY) with the difference in effectiveness (delta QALYs) between two interventions (internet intervention versus control). This expresses the difference in effects in monetary terms. Subtracting the difference in costs between the two conditions (delta costs) results in the INB (see formula 7). WTP for one QALY was set at $40,000 dollars. This value was based on WTP values in high-income countries, which are typically around $40,000 per QALY (Schwarzer et al., 2015).

Pooling incremental net benefits

As studies were expected to be heterogeneous concerning follow-up periods, included costs and sampled population, a random effects approach was maintained throughout the analyses, regardless of heterogeneity scores like the Cochran Q and I-squared. In accordance with this approach, study weights were corrected based on the DerSimonian and Laird method (DerSimonian & Laird, 1986).

Moderators

Several moderators were incorporated in the analyses to explore their influence on the overall cost-effectiveness. Specifically, subgroups were based on (1) the perspective of the economic evaluation (health care versus societal), (2) length of follow-up (12 months or longer versus shorter than 12 months) and (3) targeted mental disorder. Other considered subgroups were based on (4) presence of guidance (yes/no), (5) intensity of the guidance (self-guided, less than weekly, weekly or more than weekly), (6) type of guidance (asynchronous/delayed such as e-mail, synchronous/immediate such as chat or telephone, or a combination), (7) method of recruitment (open/mass media or clinical referral), (8) method of diagnosis for inclusion (formal diagnosis or self-reported symptoms), (9) duration of the intervention (4-8 weeks, 9-12 weeks, longer than 12 weeks or undefined/unlimited access), and (10) type of control condition (care as usual or attention control).

Heterogeneity and publication bias

To get an impression of the heterogeneity between studies I^2 and Cochran Q (formulas 12 and 13) were calculated and reported for all analyses. However, visual inspection of the forest plot and consideration of the study characteristics were leading in the identification of between-study heterogeneity. Indications of publication bias were explored with both a visual inspection of the funnel plot and the Eggers test.

Sensitivity analyses

Robustness of the results were inspected in five sensitivity analyses. Specifically, the pooled INB was calculated separately for studies with (1) high quality economic evaluations based

on the CHEC list, and (2) low risk of bias based on the Cochrane risk of bias tool. Two analyses (3 and 4) explored the impact of the value of the WTP per QALY (set at $40,000 in the main analysis), by repeating the analysis with WTP values of $20,000 and $80,000. In the last analysis (5) only studies were pooled for which the covariance could be calculated directly from author data.

Results

Figure 1 comprises the study selection flow. Table 1 and Table 2 present a detailed overview of all included studies and their outcomes. Additionally, an overview of all studies that were excluded after the full-text screening phase and the reason for exclusion can be found in the Appendix C.6. In total 6226 papers, conference abstracts and trial registrations were identified. Full texts were examined of 178 papers and finally data from 37 articles published between 1990 and March 2021 were extracted. From 200 relevant protocols, conference abstracts and trial registrations, follow up was needed for 93 records with 76 different corresponding authors. For those not needing follow up it was clear that data gathering was still ongoing or data was already published and included in the screening. The 76 corresponding authors were approached via e-mail to clarify whether data were already available to implement in our meta-analysis. One reminder was sent within two weeks if there was no response. This yielded four additional articles to include from three different authors. In total, 53 (69.7%) authors responded and for 50 of those responses (94.3%) cost-utility data were not (yet) available or authors were not willing to share data at this time.

Table 1. Characteristics of included studies.

Characteristic	N	References
Country		
United Kingdom	12	Bogosian et al. (2015) and Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Dixon et al. (2016), Duarte et al. (2017), Hollinghurst et al. (2010), Lovell et al. (2017), Morriss et al. (2019), Powell et al. (2020), Richards et al. (2020), and Wright et al. (2020)
Netherlands	8	Aardoom, Dingemans, van Ginkel, et al. (2016), Ferwerda et al. (2018), Geraedts et al. (2015), Gerhards et al. (2010), Kolovos et al. (2016), Lokman et al. (2017), Van Luenen et al. (2019), and Warmerdam et al. (2010)
Sweden	5	Holst et al. (2018), Jolstedt et al. (2018), Kraepelien et al. (2018), Lenhard et al. (2017), and Lindsäter et al. (2019)
Germany	4	Buntrock et al. (2017), Kählke et al. (2019), Nobis et al. (2018), and Röhr et al. (2021)
Australia	3	Dear et al. (2015), Moayeri et al. (2019), and Titov et al. (2015)
United States	2	Joesch et al. (2012) and Murphy et al. (2016)
Canada	1	Yan et al. (2019)
Italy	1	Hunter et al. (2017)
Spain	1	Romero-Sanchiz et al. (2017)
Targeted disorder		

Depression	16	Buntrock et al. (2017), Dixon et al. (2016), Duarte et al. (2017), Geraedts et al. (2015), Gerhards et al. (2010), Hollinghurst et al. (2010), Holst et al. (2018), Kolovos et al. (2016), Kraepelien et al. (2018), Nobis et al. (2018), Romero-Sanchiz et al. (2017), Titov et al. (2015), Van Luenen et al. (2019), Warmerdam et al. (2010), Wright et al. (2020), and Yan et al. (2019)
Anxiey	7	Dear et al. (2015), Joesch et al. (2012), Jolstedt et al. (2018), Kählke et al. (2019), Lindsäter et al. (2019), Morriss et al. (2019), and Powell et al. (2020)
Substance abuse	5	Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Hunter et al. (2017) and Murphy et al. (2016)
Depression / anxiety	5	Bogosian et al. (2015), Ferwerda et al. (2018), Lokman et al. (2017), Moayeri et al. (2019), and Richards et al. (2020)
Obsessive compulsive disorder	2	Lenhard et al. (2017) and Lovell et al. (2017)
PTSD	1	Röhr et al. (2021)
Eating disorders	1	Aardoom, Dingemans, van Ginkel, et al. (2016)
Method of diagnosis for inclusion		
Formal diagnosis	13	Buntrock et al. (2017) and Dear et al. (2015)[a], Dixon et al. (2016)[a], Hollinghurst et al. (2010), Holst et al. (2018)[a], Joesch et al. (2012), Jolstedt et al. (2018), Kolovos et al. (2016), and Lenhard et al. (2017)[a], Lindsäter et al. (2019), Lovell et al. (2017), Morriss et al. (2019), and Romero-Sanchiz et al. (2017)
Self-reported symptoms	23	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), and Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Duarte et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Gerhards et al. (2010), Hunter et al. (2017), Kählke et al. (2019), Kraepelien et al. (2018), Lokman et al. (2017), Moayeri et al. (2019), Murphy et al. (2016), Nobis et al. (2018), Powell et al. (2020), and Richards et al. (2020)[a], Röhr et al. (2021) and Titov et al. (2015)[a], Van Luenen et al. (2019), Warmerdam et al. (2010), and Wright et al. (2020)
Other	1	Yan et al. (2019) (no assessment before inclusion)
Recruitment type		
Via clinical institution	20	Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Dixon et al. (2016), Duarte et al. (2017), Ferwerda et al. (2018), Hollinghurst et al. (2010), Holst et al. (2018), Hunter et al. (2017), Joesch et al. (2012), Kolovos et al. (2016), Kraepelien et al. (2018), Lovell et al. (2017), Moayeri et al. (2019), Morriss et al. (2019), Murphy et al. (2016), Richards et al. (2020), Romero-Sanchiz et al. (2017), Van Luenen et al. (2019), Wright et al. (2020), and Yan et al. (2019)
Open or mass media recruitment	13	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Buntrock et al. (2017), Dear et al. (2015), Gerhards et al. (2010), Jolstedt et al. (2018), Kählke et al. (2019), Lindsäter et al. (2019), Nobis et al. (2018), Powell et al. (2020), Röhr et al. (2021), Titov et al. (2015), and Warmerdam et al. (2010)
Other[b]	4	Crombie et al. (2018), Geraedts et al. (2015), Lenhard et al. (2017), and Lokman et al. (2017)
Economic perspective		

Societal	22	Aardoom, Dingemans, van Ginkel, et al. (2016), Buntrock et al. (2017), and Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Dixon et al. (2016), Ferwerda et al. (2018), Geraedts et al. (2015), Gerhards et al. (2010), Holst et al. (2018), Jolstedt et al. (2018), Kählke et al. (2019), Kolovos et al. (2016), Kraepelien et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Lokman et al. (2017), Lovell et al. (2017), Nobis et al. (2018), Romero-Sanchiz et al. (2017), Van Luenen et al. (2019), and Warmerdam et al. (2010)
Healthcare	15	Bogosian et al. (2015), Dear et al. (2015), Duarte et al. (2017), Hollinghurst et al. (2010), Hunter et al. (2017), Joesch et al. (2012), Moayeri et al. (2019), Morriss et al. (2019), Murphy et al. (2016), Powell et al. (2020), Richards et al. (2020), Röhr et al. (2021), Titov et al. (2015), Wright et al. (2020), and Yan et al. (2019)
Follow-up period		
8-12 weeks	6	Dear et al. (2015), Jolstedt et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Moayeri et al. (2019), and Warmerdam et al. (2010)
4-6 months	6	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Kählke et al. (2019), Nobis et al. (2018), Röhr et al. (2021), and Van Luenen et al. (2019)
8-9 months	2	Hollinghurst et al. (2010) and Murphy et al. (2016)
12-14 months	20	Buntrock et al. (2017) and Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Dixon et al. (2016), Geraedts et al. (2015), Gerhards et al. (2010), Holst et al. (2018), Hunter et al. (2017), Kolovos et al. (2016), Kraepelien et al. (2018), Lokman et al. (2017), Lovell et al. (2017), Morriss et al. (2019), Powell et al. (2020), Richards et al. (2020), Romero-Sanchiz et al. (2017), Titov et al. (2015), Wright et al. (2020), and Yan et al. (2019)
18 months or longer	3	Duarte et al. (2017), Ferwerda et al. (2018), and Joesch et al. (2012)
Intervention duration		
Shorter than 6 weeks	2	Kolovos et al. (2016) and Röhr et al. (2021)
6-8 weeks	15	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Buntrock et al. (2017), Dear et al. (2015), Duarte et al. (2017), Geraedts et al. (2015), Gerhards et al. (2010), Kählke et al. (2019), Nobis et al. (2018), Powell et al. (2020), Richards et al. (2020), Titov et al. (2015), Van Luenen et al. (2019), Warmerdam et al. (2010), and Wright et al. (2020)
9-12 weeks	13	Crombie et al. (2018), Holst et al. (2018), Joesch et al. (2012), Jolstedt et al. (2018), Kraepelien et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Lovell et al. (2017), Moayeri et al. (2019), Morriss et al. (2019), Murphy et al. (2016), Romero-Sanchiz et al. (2017), and Yan et al. (2019)
Longer than 12 weeks	3	Dixon et al. (2016), Ferwerda et al. (2018), and Hollinghurst et al. (2010)
Undefined or unlimited access	4	Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Hunter et al. (2017) and Lokman et al. (2017)
Human guidance available		

Yes	27	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Buntrock et al. (2017), Dear et al. (2015), Dixon et al. (2016), Duarte et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Hollinghurst et al. (2010), Holst et al. (2018), Joesch et al. (2012), Jolstedt et al. (2018), Kählke et al. (2019), Kolovos et al. (2016), Kraepelien et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Lovell et al. (2017), Moayeri et al. (2019), Morriss et al. (2019), Murphy et al. (2016), Nobis et al. (2018), Richards et al. (2020), Titov et al. (2015), Van Luenen et al. (2019), Warmerdam et al. (2010), and Wright et al. (2020)
No	10	Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Gerhards et al. (2010), Hunter et al. (2017), Lokman et al. (2017), Powell et al. (2020), Röhr et al. (2021), Romero-Sanchiz et al. (2017), and Yan et al. (2019)
Guidance type		
Self-guided	10	Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Gerhards et al. (2010), Hunter et al. (2017), Lokman et al. (2017), Powell et al. (2020), Röhr et al. (2021), Romero-Sanchiz et al. (2017), and Yan et al. (2019)
Email or written support	11	Buntrock et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Jolstedt et al. (2018), Kählke et al. (2019), Kolovos et al. (2016), Kraepelien et al. (2018), Lindsäter et al. (2019), Nobis et al. (2018), Richards et al. (2020), and Warmerdam et al. (2010)
Chat support	1	Hollinghurst et al. (2010)
Telephone support	5	Dixon et al. (2016), Duarte et al. (2017), Lovell et al. (2017), Moayeri et al. (2019), and Van Luenen et al. (2019)
Video conferencing	1	Bogosian et al. (2015)
Face-to-face support	3	Joesch et al. (2012), Murphy et al. (2016), and Wright et al. (2020)
Combination	6	Aardoom, Dingemans, van Ginkel, et al. (2016), Dear et al. (2015), Holst et al. (2018), Lenhard et al. (2017), Morriss et al. (2019), and Titov et al. (2015)
Guidance frequency		
Not applicable	10	Crombie et al. (2018), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Gerhards et al. (2010), Hunter et al. (2017), Lokman et al. (2017), Powell et al. (2020), Röhr et al. (2021), Romero-Sanchiz et al. (2017), and Yan et al. (2019)
Less than weekly	3	Dixon et al. (2016), Hollinghurst et al. (2010), and Lovell et al. (2017)
weekly	21	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Buntrock et al. (2017), Dear et al. (2015), Duarte et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Holst et al. (2018), Joesch et al. (2012), Jolstedt et al. (2018), Kählke et al. (2019), Kolovos et al. (2016), Lindsäter et al. (2019), Moayeri et al. (2019), Morriss et al. (2019), Nobis et al. (2018), Richards et al. (2020), Titov et al. (2015), Van Luenen et al. (2019), Warmerdam et al. (2010), and Wright et al. (2020)
More than weekly	3	Kraepelien et al. (2018), Lenhard et al. (2017), and Murphy et al. (2016)
Control condition type		

Care as usual[c]	32	Aardoom, Dingemans, van Ginkel, et al. (2016), Bogosian et al. (2015), Buntrock et al. (2017), and Dear et al. (2015), Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk], Dixon et al. (2016), Duarte et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Gerhards et al. (2010), Hollinghurst et al. (2010), Holst et al. (2018), Hunter et al. (2017), Joesch et al. (2012), Kählke et al. (2019), Kolovos et al. (2016), Kraepelien et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Lokman et al. (2017), Lovell et al. (2017), Morriss et al. (2019), Murphy et al. (2016), Nobis et al. (2018), Powell et al. (2020), Richards et al. (2020), Röhr et al. (2021), Romero-Sanchiz et al. (2017), Titov et al. (2015), Warmerdam et al. (2010), and Yan et al. (2019)
Attention control	5	Crombie et al. (2018), Jolstedt et al. (2018), Moayeri et al. (2019), Van Luenen et al. (2019), and Wright et al. (2020)
Mode of delivery		
Website[d]	28	Aardoom, Dingemans, van Ginkel, et al. (2016), Buntrock et al. (2017), Dear et al. (2015), Duarte et al. (2017), Ferwerda et al. (2018), Geraedts et al. (2015), Gerhards et al. (2010), Holst et al. (2018), Hunter et al. (2017), Joesch et al. (2012), Jolstedt et al. (2018), Kählke et al. (2019), Kolovos et al. (2016), Kraepelien et al. (2018), Lenhard et al. (2017), Lindsäter et al. (2019), Lokman et al. (2017), Lovell et al. (2017), Murphy et al. (2016), Nobis et al. (2018), Powell et al. (2020), Richards et al. (2020), Romero-Sanchiz et al. (2017), Titov et al. (2015), Van Luenen et al. (2019), Warmerdam et al. (2010), Wright et al. (2020), and Yan et al. (2019)
Telephone or videoconferencing	4	Bogosian et al. (2015), Dixon et al. (2016), Moayeri et al. (2019), and Morriss et al. (2019)
Chat	1	Hollinghurst et al. (2010)
Text messaging	1	Crombie et al. (2018)
App	1	Röhr et al. (2021)
Game (app or website)	2	Deluca et al. (2020) [High risk], Deluca et al. (2020) [Low risk]

[a] These studies had inclusion criteria based on both (the absence of) a formal diagnosis and self-reported symptoms or a formal diagnosis was conducted after a self-reported symptom-based inclusion
[b] Could involve a mixture of recruitment strategies, targeting a specific group of people, or recruitment via companies
[c] Could involve a waitlist or do-nothing approach (where often participants were allowed to use other forms of treatment during the study period) or a one-time informational session or flyer
[d] Interventions consisted of (often weekly) modules with (cognitive-behavioral) exercises ($n = 27$) or web-based self-monitoring ($n = 1$)

Table 2. Sample and outcomes of included studies.

Author (year of publication)	Sample size	Mean age (SD)	% Female	Delta QALY (SE)	Delta costs in US dollars (SE)	INB in US dollars (VAR)
Aardoom, Dingemans, van Ginkel, et al. (2016)	354	24.2 (7.7)	99.0	$< .01(.01)$	-660 (433)	668 (489207)
Bogosian et al. (2015)	40	52.7 (9.5)	55.0	$-.01(.02)$	-3216 (2056)	2976 (4688465)

Buntrock et al. (2017)	406	45.0 (11.9)	73.9	$.01(.02)$	169 (179)	232 (884732)
Crombie et al. (2018)	825	35.0 (miss)	0.0	$-.01(.02)$	488 (357)	-728 (657104)
Dear et al. (2015)	70	65.5 (5.13)	60.0	$.01(.03)$	61 (19)	339 (1847789)
Deluca et al. (2020) [High risk]	756	16.1 (0.9)	50.2	$-.01(.01)$	547 (722)	-930 (737655)
Deluca et al. (2020) [Low risk]	883	15.2 (1.0)	51.7	$-.01(.01)$	639 (668)	-1058 (693923)
Dixon et al. (2016)	609	49.6 (12.8)	68.5	$< .01(.01)$	2805 (144)	-196 (281174)
Duarte et al. (2017)	691	39.9 (12.7)	67.0	$-.04(.04)$	171 (145)	-1911 (2622289)
Ferwerda et al. (2018)	133	56.4 (10.0)	64.9	$.06(.03)$	5035 (3112)	-2675 (10740997)
Geraedts et al. (2015)	231	43.4 (9.2)	62.3	$< .01(.03)$	-889 (2962)	889 (10926230)
Gerhards et al. (2010)	303	44.9 (11.6)	43.2	$-.01(.02)$	66 (1901)	-466 (4748077)
Hollinghurst et al. (2010)	297	34.9 (11.6)	68.0	$.03(.02)$	778 (106)	302 (730265)
Holst et al. (2018)	90	38.6 (11.7)	77.8	$-.05(.03)$	-41 (58)	-1718 (6932833)
Hunter et al. (2017)	763	Median 49 (IQR 35-61)	38.5	$< .01(< .01)$	3 (4)	21 (36931)
Joesch et al. (2012)	690	45.1 (13.2)	71.7	$.05(.04)$	257 (523)	1743 (3150182)
Jolstedt et al. (2018)	131	10.0 (1.3)	53.4	$< .01(.09)$	-347 (6)	42 (15261387)
Kählke et al. (2019)	264	43.3 (10.2)	73.1	$.01(.00)$	-374 (690)	670 (706242)
Kolovos et al. (2016)	269	38.0 (11.4)	53.9	$.03(0.15)$	571 (310)	138 (37849735)
Kraepelien et al. (2018)	945	43.0 (12.2)	72.9	$.01(.03)$	-46 (496)	-1519 (4559865)
Lenhard et al. (2017)	67	14.6 (1.7)	46.0	$< .01(< .01)$	19 (5)	121 (269039)
Lindsäter et al. (2019)	100	46.2 (8.8)	85.0	$< .01(.02)$	-71 (317)	243 (1068351)
Lokman et al. (2017)	220	44.2 (9.9)	59.1	$.02(.02)$	-5000 (2740)	5840 (9751347)
Lovell et al. (2017)	473	Median 33 (range 18-77)	60.3	$.01(.01)$	-25 (251)	465 (430364)

Moayeri et al. (2019)	110	68.1 (8.8)	65.5	$-.01(.01)$	-270 (23)	-54 (197204)
Morriss et al. (2019)	156	Median 32 (range 19-82)	69.2	$.07(.07)$	-1419 (1299)	4219 (23565882)
Murphy et al. (2016)	507	34.9 (10.9)	37.9	$< .01(.01)$	147 (200)	-187 (170067)
Nobis et al. (2018)	256	51.0 (12.0)	62.9	$.01(.01)$	1001 (999)	278 (1457706)
Powell et al. (2020)	2116	37.2 (13.8)	80.2	$.01(.01)$	-87 (82)	647 (51304)
Richards et al. (2020)	361	Median 29 (IQR = 18)	71.5	$.02(.01)$	125 (65)	707 (292998)
Röhr et al. (2021)	133	33.3 (11.2)	38.3	$< .01(.01)$	-124 (139)	-36 (64494)
Romero-Sanchiz et al. (2017)	296	42.9 (10.3)	75.7	$.08(.03)$	-265 (459)	3526 (1329167)
Titov et al. (2015)	54	65.3 (3.0)	70.4	$.01(< .01)$	33 (25)	447 (30236)
Van Luenen et al. (2019)	188	46.3 (10.6)	11.7	$.01(< .01)$	13 (171)	1180 (690364)
Warmerdam et al. (2010)	263	45.0 (12.1)	71.1	$.01(.01)$	276 (613)	124 (442031)
Wright et al. (2020)	139	15.0 (1.4)	64.0	$.03(.06)$	-20 (161)	1180 (5895741)
Yan et al. (2019)	1407	47.1 (17.0)	73.0	$.01(.01)$	-140 (36)	356 (1491)

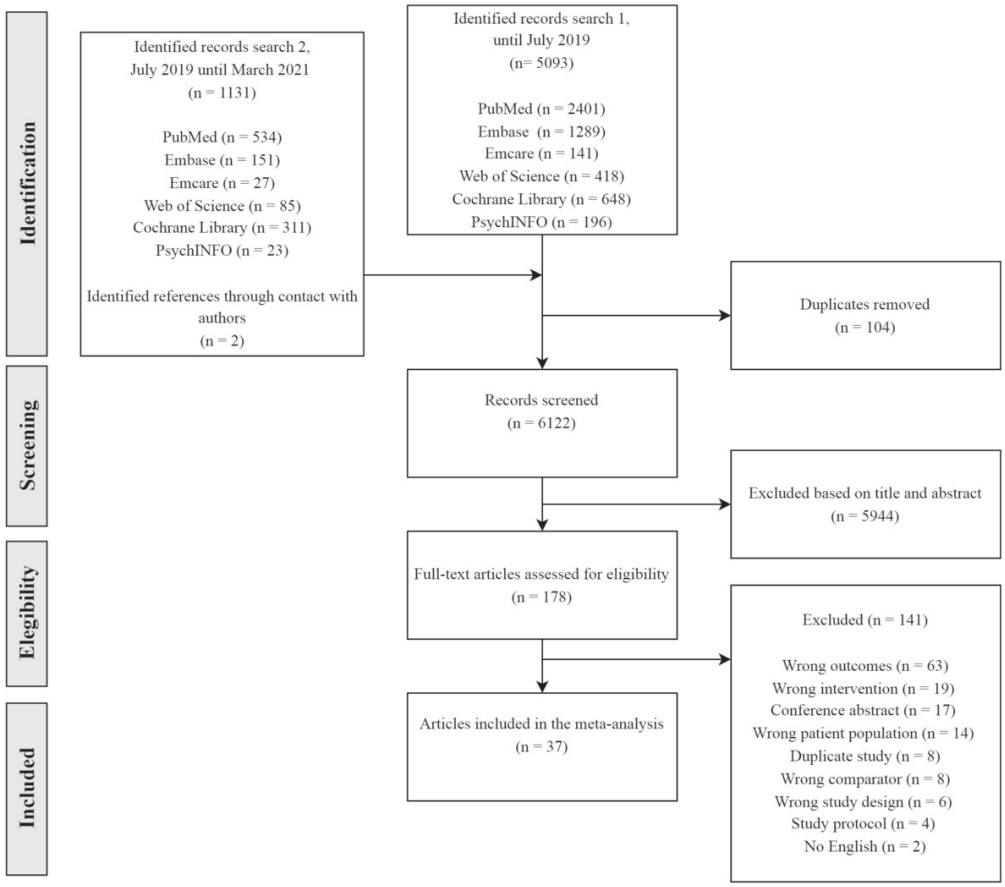
Miss=missing; IQR=Interquartile range; SD=Standard deviation; SE=standard error; VAR=variance.

Data preparation

For 7 studies no measure of spread of delta QALY or delta costs could, directly or indirectly, be deduced from the article. Data from one study were available within our research team (Aardoom, Dingemans, van Ginkel, et al., 2016). After inquiry with the authors of the other six studies (5 of which responded) data of only three comparisons were still missing. Two were solved by using Webplot Digitizer on the presented cost-effectiveness plane and the other by mean imputation based on two comparisons within the same study.

All authors of included studies were contacted to provide information on the covariance between delta QALY and delta costs. Authors responded for 31 (83.8%) of the 37 included studies, but not all authors were able to provide information on the covariance. Specifically, for 12 of the studies the covariance could be based on data from authors. The mean Pearson correlation between delta QALY and delta costs based on these 12 studies was $r = -.12$ ($SD = .16$). For the remaining 25 studies, covariances were calculated using Webplot Digitizer ($n = 13$) or using the estimated mean correlation calculated earlier ($n = 12$).

Figure 1. Study selection flow.



Characteristics of included studies

In total, the 37 included studies (Table 1) recruited 15,596 participants, ranging between 40 to 2,116. Some study conditions were irrelevant for this meta-analysis (e.g., an active intervention without internet component), so the main analysis was based on 14,946 participants. Intention-to-treat analyses were conducted in most studies ($n = 32$). Mental disorders that were targeted were depression ($n = 16$), anxiety ($n = 7$), alcohol or substance abuse ($n = 5$), depression and anxiety simultaneously ($n = 5$), obsessive compulsive disorder ($n = 2$), post-traumatic stress disorder ($n = 1$), and eating disorders ($n = 1$). Experimental interventions were mostly cognitive-behavioral based modules or websites that participants could engage with ($n = 28$). Other interventions consisted of teleconferencing ($n = 2$), text messaging ($n = 3$), a web-based game ($n = 2$) and telephone support ($n = 2$). In 27 studies some form of guidance within the intervention was available, whereas the intervention was self-guided in 10 studies. Guidance consisted of written feedback ($n = 11$), telephone calls ($n = 5$), face-to-face, including teleconferencing ($n = 4$), chat ($n = 1$), or a combination of these ($n = 6$). Control conditions of studies included waiting list or care as usual conditions ($n = 32$) and psychological placebo or attention control conditions ($n = 5$). Follow-up periods ranged from 8 weeks to 2 years, with most studies ($n = 20$) maintaining a 12 month follow-up period.

Quality of life

Questionnaires used to calculate QALYs were EQ-5D ($n = 32$), KIDSCREEN-10 ($n = 1$), SF-6D or SF-36 ($n = 3$) and the Australian quality of life instrument ($n = 1$). The pooled difference in effectiveness (intervention QALY gains minus control QALY gains) for included studies was .004 QALY ($SE = .002$), Hedges' $g = 0.05$ (95% CI 0.01; 0.10, $P = .016$). While the difference was statistically significant, likely because of the large sample size, the size of the difference was deemed negligible.

Costs

Main questionnaires used to measure healthcare use and costs in included studies were the Treatment Inventory of Costs in Psychiatric Patients ($n = 14$) and Client Service Receipt Inventory ($n = 7$), but other or self-made questionnaires, medical records and diaries were also used. In total, 15 studies reported costs from a healthcare perspective and 22 presented a societal perspective. The pooled difference in costs (intervention costs minus control costs) when studies with a healthcare and societal perspective were taken together was \$49 ($SE = 40$), Hedges' $g < 0.01$ (95% CI -0.10; 0.84, $P = .96$). Considering the uncertainty in measurements of costs, the small difference indicates that internet interventions were equally expensive as control conditions. Results for studies with different economic perspectives were similar, with no difference in costs for studies with a healthcare (\$40, $SE = 44$), Hedges' $g = -0.03$ (95% CI -0.21; 0.16, $P = .78$), and societal perspective (\$158, $SE = 76$), Hedges' $g = 0.03$ (95% CI -0.01; 0.10, $P = .15$).

Cost-effectiveness

Visual inspection of the individual INBs and their CIs (Figure 2) indicated substantial heterogeneity between the 37 included studies. Statistical measures of heterogeneity suggested otherwise, Cochran $Q(36) = 37.12$, $P = .42$, $I^2 = 3.0\%$ (95% CI 0.0%; 42.8%), but are difficult to interpret because of the large 95% CI of the I^2 and considerable within-study uncertainty. In other words, the between-study heterogeneity seemed to be overshadowed by the large within-study heterogeneity. Therefore, still a random-effects model was preferred over a fixed-effect model to pool INBs. The INB was positive (more favorable balance of costs and effects in the internet intervention compared to the control condition) in 25 of the included RCTs. Furthermore, at a WTP of $40,000 per QALY, the pooled INB was $255 (95% CI $91; $419, $P = .002$). The results suggest that internet interventions are slightly more cost-effectiveness compared to a do-nothing or care-as-usual approach.

Moderator analyses

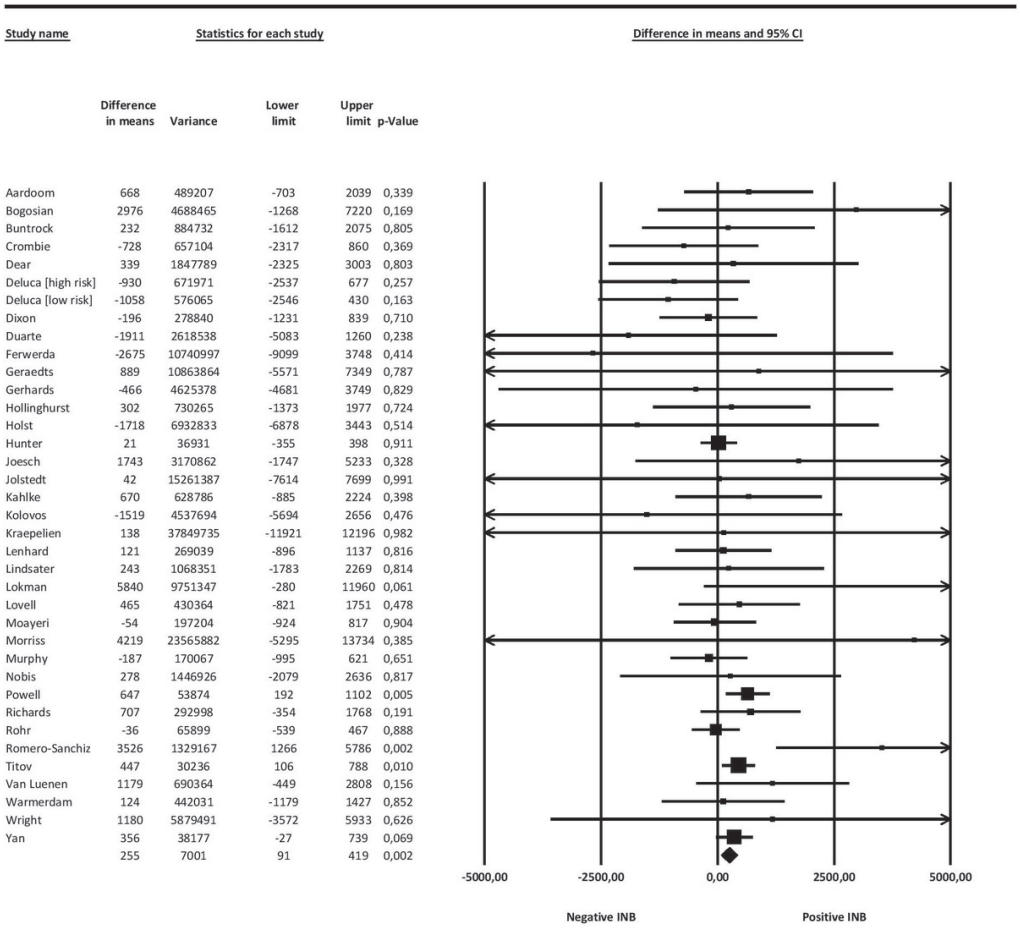
Pooled INBs were also calculated for subgroups based on the ten moderator variables. Outcomes for subgroups based on perspective, length of follow-up and targeted mental disorder are presented in the text and results for all moderator (including presence of guidance, intensity of guidance, type of guidance, recruitment strategy, diagnosis for inclusion, intervention duration, and control condition type) analyses can be found in the Appendix C.7.

Pooling studies based on economic perspective influenced the results. Specifically, looking at studies with a healthcare perspective separately ($n = 15$), the pooled INB was $280 (95% CI $109; $451, $P = .001$). For studies with a societal perspective ($n = 22$), the pooled INB was substantially lower at $161 (95% CI $-247; $569, $P = .44$). This suggests that cost-effectiveness of internet interventions compared to control conditions cannot be assumed when maintaining a societal perspective.

Studies with a short (shorter than 12 months) follow-up ($n = 14$) had a pooled INB of $112 (95% CI $-194; $418, $P = .47$) and studies with a long (12 months or longer) follow-up ($n = 23$) had a pooled INB of $270 (95% CI $-14; $554, $P = .063$). For RCTs with a long follow-up, statistical significance was likely not attained because the sample size decreased compared to the main analysis. Nevertheless, the studies with a long follow-up, which are usually better able to capture all relevant costs and effects than those with a short time horizon (Basu & Maciejewski, 2019; Neumann et al., 2016), had a pooled estimate comparable to the estimate of all studies taken together. This strengthens the idea that internet interventions are likely to be cost-effective compared to control conditions on the long term.

Concerning targeted mental disorders, a significant positive INB was found for internet interventions targeting anxiety ($n = 7$), $644 (95% CI $227; $1062, $P = .002$), and depression ($n = 16$), $387 (95% CI $156; $618, $P = .001$). Similarly, the INB of studies with internet interventions targeting depression and anxiety simultaneously (n=5), pooled INB of $580 (95% CI $-584; $1744, $P = .33$), and obsessive compulsive disorder (n=2), pooled INB of $253 (95% CI $-544; $1051, $P = .53$) was positive. The size of the INB for these two groups also indicated that cost-effectiveness compared to control conditions is likely, but statistical significance was not attained. Internet interventions were unlikely to be cost-effective

Figure 2. Forest plot of incremental net benefits (INB) and the pooled estimate according to a random effects model.



Note. Difference in means indicates the incremental net benefit. The last row indicates the pooled incremental net benefit according to a random effects model.

when targeting alcohol or substance abuse (n=5), pooled INB of $-129 (95% CI $-448; $191, $P = .43$). It must be noted that of the two studies regarding obsessive compulsive disorder one targeted children, further obscuring interpretability for this subgroup. Only one study was available for eating disorders and posttraumatic stress disorder, rendering it impossible to pool.

Risk of bias and quality of economic evaluation

Table 3 comprises details on risk of bias and Table 4 on quality ratings of the economic analysis for the included studies. The overall risk of bias of the included RCTs was considered high, with low risk of bias for 8, medium risk of bias for 8 and high risk of bias for 21 studies. The economic appraisal of the included studies, based on the CHEC list, suggested moderate quality, with 24 studies receiving a high and 13 a low quality rating. Level of agreement between raters was considered low for the risk of bias assessment (55% agreement) and high for the CHEC list ratings (89% agreement).

Publication bias

A funnel plot of included studies is presented in Figure 3. Visual inspection of the plot seemed to suggest some evidence for publication bias. Egger's test did not indicate asymmetry in the funnel plot ($z = -.026$, one-tailed $P = .80$).

Sensitivity analyses

Results of all sensitivity analyses can be found in the Appendix C.8. First, the main analysis was repeated for studies with a high quality rating on the CHEC list ($n = 24$), resulting in a pooled INB of $253 (95% CI $43; $463, $P = .018$). This suggests cost-effectiveness of internet interventions was maintained when looking at high quality studies alone and that results from the main analysis were not dependent on low quality studies. Secondly, the pooled INB for studies with a low risk of bias ($n = 8$) was $244 (95% CI $-555; $1042, $P = .55$). A similar result was found when medium risk of bias was considered as low risk of bias ($n = 16$), with a pooled INB of $216 (95% CI $-182; $615, $P = .29$). The pooled INB for RCTs with low risk of bias was comparable to that of the main analysis, suggesting that the pooled result of all 37 studies was not critically biased. However, for high risk of bias studies the pooled INB was slightly higher, $272 (95% CI $68; $475, $P = .009$). This suggests a small overestimation in the overall pooled estimate. Thirdly, when one QALY was valued at $20,000 instead of $40,000 the pooled INB of the 37 studies was $145 (95% CI $56; $234, $P = .001$). At a WTP of $80,000 for one QALY the pooled INB was $431 (95% CI $115; $747, $P = .008$). The two analyses show that if society's WTP for one additional QALY for an individual is lower ($20,000) or higher ($80,000) than $40,000, cost-effectiveness of internet interventions compared to care as usual is still expected. Finally, the pooled INB of studies for which the covariance could be calculated directly from author data ($n = 12$) was $264 (95% CI $-167; $694, $P = .23$). This was similar to the pooled estimate of all 37 studies, indicating that the way the covariance was calculated did not greatly influence the results of the meta-analysis.

Deviations from the protocol

An individual participant data meta-analysis was not achievable, so moderator analyses based on age, gender and symptom severity were not possible or feasible. Additionally, a moderator analysis with subgroups based on whether the study was conducted by developers of the interventions or not was planned, but this information proved too difficult to find for many of the studies.

Table 3. Cochrane risk of bias table of included studies.

Author (year of publication)	1	2	3	4	5	6	7	Overall Risk of Bias
Aardoom, Dingemans, van Ginkel, et al. (2016)	+	+	-	+	+	+	+	Low
Bogosian et al. (2015)	+	+	-	+	+	?	+	Low
Buntrock et al. (2017)	+	+	-	+	+	+	+	Low
Crombie et al. (2018)	+	+	+	+	+	+	+	Low
Dear et al. (2015)	+	+	-	-	?	?	+	High
Deluca et al. (2020) [High risk]	+	+	-	+	-	+	+	High
Deluca et al. (2020) [Low risk]	+	+	-	+	-	+	+	High
Dixon et al. (2016)	+	+	-	+	+	+	?	Medium
Duarte et al. (2017)	+	+	-	+	?	+	-	High
Ferwerda et al. (2018)	+	+	-	+	-	?	+	High
Geraedts et al. (2015)	+	+	-	+	+	?	+	Low
Gerhards et al. (2010)	+	+	-	+	-	?	+	High
Hollinghurst et al. (2010)	+	+	-	?	-	?	-	High
Holst et al. (2018)	+	+	-	-	?	+	?	High
Hunter et al. (2017)	+	+	-	-	+	+	-	High
Joesch et al. (2012)	+	+	-	+	-	+	+	High
Jolstedt et al. (2018)	+	+	-	+	+	+	+	Low
Kählke et al. (2019)	+	+	-	?	+	+	+	Medium
Kolovos et al. (2016)	+	?	-	+	+	+	+	Medium
Kraepelien et al. (2018)	+	+	-	+	+	?	+	Low
Lenhard et al. (2017)	+	+	-	+	+	+	-	High
Lindsäter et al. (2019)	+	+	-	+	+	+	+	Low
Lokman et al. (2017)	?	+	-	+	-	?	-	High
Lovell et al. (2017)	+	+	-	?	+	+	-	High
Moayeri et al. (2019)	+	+	-	+	+	?	?	Medium
Morriss et al. (2019)	+	+	-	+	+	?	?	Medium
Murphy et al. (2016)	+	+	-	?	+	?	-	High
Nobis et al. (2018)	+	+	-	+	+	-	+	High
Powell et al. (2020)	+	+	-	+	-	?	+	High
Richards et al. (2020)	+	-	-	-	-	+	+	High
Röhr et al. (2021)	+	+	-	?	+	+	+	Medium
Romero-Sanchiz et al. (2017)	+	+	-	+	+	?	?	Medium
Titov et al. (2015)	+	+	-	-	?	?	?	High

Van Luenen et al. (2019)	+	+	-	?	+	?	+	Medium
Warmerdam et al. (2010)	+	?	-	+	+	?	-	High
Wright et al. (2020)	+	+	-	+	-	?	?	High
Yan et al. (2019)	-	-	-	?	-	+	-	High

1=Random sequence generation, 2=Allocation concealment, 3=Blinding of participants and personnel, 4=Incomplete outcome data, 5=Blinding of outcome assessors, 6=Selective reporting, 7=Other bias
+=item scored as low risk of bias.
-=item scored as high risk of bias.
?=item scored as unclear risk of bias.

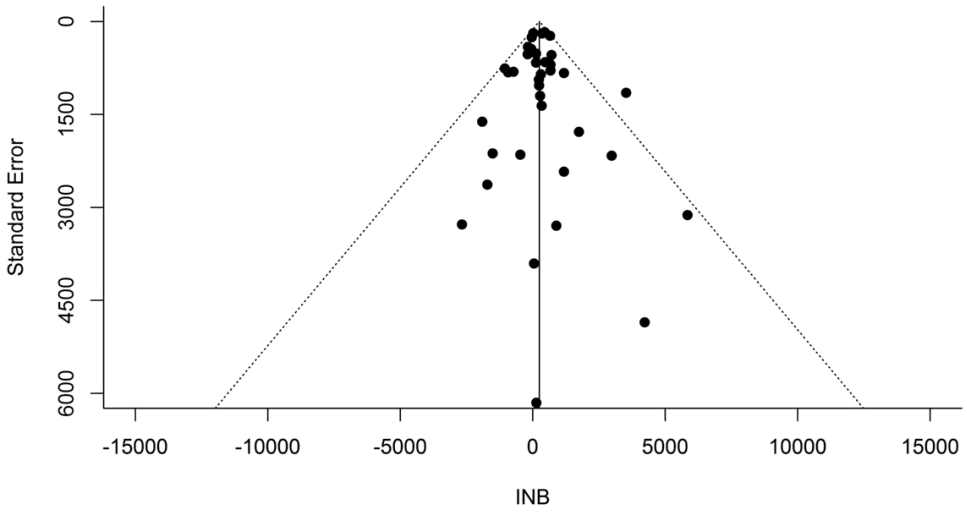
Table 4. Quality of the economic evaluation of included studies using the CHEC list.

Author (year of publication)	1	2	3	4	5	6	7[a]	8[a]	9[a]	10[a]	11[a]	12[a]	13[a]	14[a]	15	16	17	18	19	Sum score	Quality
Aardoom, Dingemans, van Ginkel, et al. (2016)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	17/19	H
Bogosian et al. (2015)	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	+	-	+	-	14/19	H
Buntrock et al. (2017)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	19/19	H
Crombie et al. (2018)	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	18/19	L
Dear et al. (2015)	+	+	+	+	-	-	-	+	+	+	+	+	+	+	-	+	+	+	+	14/19	L
Deluca et al. (2020) [High risk]	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	15/19	L
Deluca et al. (2020) [Low risk]	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	-	+	-	15/19	L
Dixon et al. (2016)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	17/19	H
Duarte et al. (2017)	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	17/19	H
Ferwerda et al. (2018)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	17/19	H
Geraedts et al. (2015)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	18/19	H
Gerhards et al. (2010)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	18/19	H
Hollinghurst et al. (2010)	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	+	+	16/19	H
Holst et al. (2018)	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	+	-	16/19	L
Hunter et al. (2017)	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	-	+	-	15/19	H
Joesch et al. (2012)	+	+	-	+	+	-	-	+	+	+	+	+	+	-	-	+	+	+	-	13/19	L
Jolstedt et al. (2018)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	+	+	+	+	17/19	H
Kählke et al. (2019)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	18/19	H
Kolovos et al. (2016)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	17/19	H
Kraepelien et al. (2018)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	18/19	H
Lenhard et al. (2017)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	18/19	H

Lindsäter et al. (2019)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	17/19	H
Lokman et al. (2017)	+	+	-	+	+	+	+	+	+	+	+	-	-	+	+	-	+	+	-	14/19	L
Lovell et al. (2017)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	17/19	H
Moayeri et al. (2019)	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	16/19	H
Morriss et al. (2019)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	17/19	H
Murphy et al. (2016)	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	+	+	-	16/19	L
Nobis et al. (2018)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	16/19	H
Powell et al. (2020)	+	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	-	14/19	L
Richards et al. (2020)	+	+	+	-	+	-	+	+	+	+	-	+	+	+	+	+	+	+	-	15/19	L
Röhr et al. (2021)	+	+	+	+	-	-	+	+	+	+	+	+	-	+	+	+	+	+	-	15/19	L
Romero-Sanchiz et al. (2017)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	17/19	H
Titov et al. (2015)	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	+	+	+	-	15/19	H
Van Luenen et al. (2019)	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	16/19	L
Warmerdam et al. (2010)	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	17/19	H
Wright et al. (2020)	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	-	16/19	H
Yan et al. (2019)	+	+	+	-	+	-	+	+	-	+	-	+	+	+	+	+	-	+	-	13/19	L

H=High overall quality; L=Low overall quality.
1=Is the study population clearly described?
2=Are competing alternatives clearly described?
3=Is a well-defined research question posed in answerable form?
4=Is the economic study design appropriate to the stated objective?
5=Is the chosen time horizon appropriate to include relevant costs and consequences?
6=Is the actual perspective chosen appropriate?
7=Are all important and relevant costs for each alternative identified?
8=Are all costs measured appropriately in physical units?
9=Are costs valued appropriately?
10=Are all important and relevant outcomes for each alternative identified?
11=Are all outcomes measured appropriately?
12=Are outcomes valued appropriately?
13=Is an incremental analysis of costs and outcomes of alternatives performed?
14=Are all future costs and outcomes discounted appropriately?
15=Are all important variables, whose values are uncertain, appropriately subjected to sensitivity analysis?
16=Do the conclusions follow from the data reported?
17=Does the study discuss the generalizability of the results to other settings and patient/client groups?
18=Does the article indicate that there is no potential conflict of interest of study researcher(s) and funder(s)?
19=Are ethical and distributional issues discussed appropriately?
[a] necessary item for a high quality score.
+=item scored as sufficient/high quality.
-=item scored as insufficient/low quality.

Figure 3. Funnel plot of standard error by incremental net benefit (INB) for inspecting publication bias.



Discussion

This systematic review and meta-analysis aimed to research the pooled evidence of the cost-effectiveness of internet interventions for mental disorders compared to control conditions. Results indicated that internet interventions were negligibly more effective in terms of QALY gains and equally costly compared to control conditions, but might be cost-effective. Internet interventions had an INB of $255 (95% CI $91; $419) compared to control conditions when society is willing to pay $40,000 for one QALY improvement for an individual. INBs were still positive when WTP for one additional QALY was lower ($20,000) or higher ($80,000). The results suggest that internet interventions for mental disorders are likely to be cost-effective compared to a do-nothing approach, especially when they target depression or anxiety and when some form of guidance is added to the intervention. This is especially interesting considering that internet interventions are now frequently and successfully being implemented (Titov et al., 2018) and can be used to serve populations that need mental health care but do not yet receive it (Ebert et al., 2018). Financing such scalable and, in some cases, anonymous interventions comes with difficulties, as they often do not fit in the funding possibilities for traditional treatments. Nevertheless, findings from this meta-analysis indicate that doing so might ultimately help to reduce the global burden of mental disorders.

Moderator analyses revealed that cost-effectiveness of internet interventions compared to control conditions was maintained for studies with a health care perspective, but not for those with a societal perspective. An explanation could be that indirect costs included in a societal perspective are usually higher and measured with more uncertainty than direct health care costs. Indeed, studies with a health care perspective included in the analyses had both a smaller cost range and smaller pooled standard error compared to studies with a societal perspective. A relatively small difference in health care costs might then be obscured by large indirect costs. This reflects a larger problem that sample size calculations for RCTs are almost exclusively based on detecting differences in disorder-specific effectiveness while neglecting QALYs and costs, rendering the trial unlikely to be adequately powered for an economic evaluation (Hollingworth et al., 2013). Furthermore, compared to control conditions, internet interventions were likely to be cost-effective for symptoms of anxiety, depression and obsessive compulsive disorder. This was not the case for alcohol and substance abuse. The finding that internet interventions for alcohol and substance abuse were not found to be cost-effective compared to controls, might be due to a lack of power in that subgroup. Alternatively, it could relate to the absence of guidance. Indeed, guided interventions were found to be cost-effective compared to care as usual whereas self-guided ones were not, and none of the alcohol and substance abuse studies incorporated guidance. Adding guidance to an internet intervention, then, might have a positive effect on cost-effectiveness. However, self-guided interventions have less functions than guided ones and do not compensate for this by higher levels of technical sophistication (i.e., responsiveness), at least for those targeting depression (Burger et al., 2020), perhaps partly explaining the positive influence of guidance in internet interventions. Another result was that studies with a follow-up period of 12 months or longer generally showed internet interventions for mental disorders to be efficient compared to a control condition, while this was not the case for studies with shorter follow-up periods. Possibly, some factors that influence the efficiency of internet interventions become apparent after a certain time only. For example, effects of an intervention concerning health

care visits or work productivity may take a while to attain. This substantiates the idea that follow-up periods of at least 12 months are important when conducting an economic evaluation of such interventions (Basu & Maciejewski, 2019; Neumann et al., 2016). Several other findings of the subgroup analyses are worth mentioning. First, recruiting participants through (social) media rather than by clinical referral was more likely to yield a positive INB. Speculatively, studies that used an open recruitment system included participants that were less severely ill or more strongly motivated for this type of interventions, hence benefitting more. Comparing study subgroups based on severity of symptoms, motivation to change, or related participant characteristics directly was not feasible for this meta-analysis, but could be an interesting avenue for future research. Second, shorter interventions were likely to be efficient compared to controls, whereas this was not true for longer interventions. It might be that longer internet interventions cost more, but do not perform better compared to shorter interventions in terms of QALYs or health care or societal costs. Third, cost-effectiveness of internet interventions for mental disorders was not likely when control conditions included an active component (i.e., attention control). Accordingly, activating participants seems to be an important ingredient for an efficient intervention, while the content may be of lesser importance. Moderator analyses should be considered as exploratory, because of the low number of studies in some subgroups.

QALYs and mental health interventions

Interestingly, internet interventions were found to be only marginally more effective than control conditions in terms of QALY gain. It is surprising that internet interventions produced practically the same amount of QALYs as a do-nothing approach, given the substantial evidence for the effectiveness of such interventions (Andersson et al., 2019). However, it is likely that the internet interventions were more effective in terms of symptom reduction and other areas of well-being, but these improvements were not captured well by the generic health-related quality-of-life measures (e.g., EQ-5D and SF-36) used in economic evaluations. Such instruments capture only a selective amount of domains of quality of life and employ an almost exclusive focus on people's current functional abilities with little emphasis on coping capabilities and resources (Pietersma et al., 2013). Consequently, they might not be suitable in contexts outside of health, such as chronic illness, elderly care, general well-being and mental illness (Mitchell et al., 2017). Future economic analyses of internet interventions for mental disorders should consider using other instruments, such as the ICECAP-A (Al-Janabi et al., 2012), which measures well-being beyond (physical) health, to complement generic health questionnaires (Keeley et al., 2016; National Institute for Health and Care Excellence, 2016).

Meta-analyses on cost-effectiveness data

To our knowledge, meta-analyses on cost-effectiveness studies have only been attempted in five studies in different fields of medicine by one other research team (Bagepally et al., 2020; Bagepally et al., 2019; Chaiyakittisopon et al., 2021; Haider et al., 2019; Noparatayaporn et al., 2021), but not yet in the area of mental health interventions. The theoretical method by Crespo et al. (2014) has been practically applied and explained by Bagepally et al. (2020).

The current study built on this method by improving the precision of estimating covariances between delta costs and delta QALYs. In general, many data are necessary to pool cost-effectiveness outcomes in a meta-analysis, some of which are often not reported. Enhancing quality of economic evaluations by clearly reporting costs, QALYs and indicators of spread (e.g., SD or SE) for all studied conditions or open availability of study data makes conducting meta-analyses on cost-effectiveness data more feasible. Besides these practical challenges, the unavoidable and substantial heterogeneity between cost-effectiveness studies alongside RCTs might make statistically pooling outcomes undesirable (Shields & Elvidge, 2020). Therefore, careful planning and cautious interpretation of the findings are warranted when considering a meta-analysis on cost-effectiveness data. Nevertheless, bearing the limitations in mind, several considerations corroborated the choice for pooling cost-effectiveness data. First, by adhering to clear predefined inclusion and exclusion criteria throughout the search procedure, the included studies were comparable across many characteristics. For example, the interventions were often similar in terms of content, mode of delivery, duration, and frequency. Additionally, in the majority of studies the control conditions consisted of a do-nothing approach (i.e., participants did not receive the internet intervention, but were allowed to keep on receiving the care they got before the study). Relatedly, 21 of the included studies used the TiC-P or CSRI to measure service use, suggesting that these studies considered the same costs, and almost all studies used the EQ-5D for calculating QALYs. Second, modeling studies were excluded to avoid additional variation between methods. Third, characteristics that were thought to have a large influence on the outcomes were tested in moderator and sensitivity analyses. Overall, these analyses were in line with the overall pooled estimate, further substantiating the robustness of the finding that internet interventions are likely to be cost-effective compared to control conditions. Lastly, the variance of the INB of individual studies was often large. Consequently, a meta-analysis was important, since it offered insight beyond individual studies, which are rarely powered to detect differences in QALYs and costs (Hollingworth et al., 2013).

Limitations

While the main, moderator and sensitivity analyses point in the direction of internet interventions being cost-effective compared to control conditions, findings should be interpreted with considerable caution and are difficult to generalize beyond the current sample of included studies. A first limitation is that the studies were heterogeneous in terms of included costs and follow-up duration, which warrants careful interpretation of the overall pooled result. Second, variances of the cost-effectiveness outcomes were large to such an extent that they overshadowed differences between studies, making it hard to understand and quantify between-study heterogeneity. Third, publication bias possibly shifted results in favor of internet interventions, because economic analyses are often a last step in effectiveness research and some results might never get published. For example, some contacted authors who published a protocol mentioning an economic analyses replied that such an analysis was ultimately not attempted, since the internet intervention was not found to be effective. To counter this problem, researchers are encouraged to designate in their RCT study protocol whether an economic analysis will be attempted and perform it regardless of results on effectiveness. A fourth reason to interpret the results with caution is that risk of bias in

most RCTs was high, possibly leading to a slight overestimation of the efficiency of internet interventions compared to control conditions. However, when quality was assessed with the CHEC list, which is arguably more suitable in the case of economic evaluations, high quality studies showed a positive INB whereas low quality studies did not. This confirms the idea that internet interventions might be cost-effective compared to controls only when the economic evaluation is of adequate quality. The subjectively chosen, though theory driven, cut-off points for the risk of bias and CHEC list should be taken into account when considering the importance of these sensitivity analyses. Indeed, studies designated as high quality based on the CHEC might not have had an appropriate time horizon or economic perspective.

Implications and future directions

The current study gives an overview of published articles on the cost-effectiveness of internet interventions for mental disorders and their findings and provides insights for future research and implementation steps. First, it must be noted that the results should be replicated in other meta-analyses before clinical and policy implications can be stated reliably. To accomplish this, standardization of economic evaluations in the area of mental health, as has been done in other sectors (Dirksen & Evers, 2016; Hiligsmann et al., 2019), would be helpful. Consequently, included studies in similar meta-analyses would be more homogeneous and easier to compare. Nevertheless, the pooled estimates of the main and subgroup analyses complement the results of the individual studies. They suggest that policy makers, insurance companies and subsidy providers should invest in internet interventions for mental disorders, as they appear to be an efficient way of improving quality of life compared to not offering them. For clinical practice, often facing financial pressure and waiting lists, this might involve a transition where internet interventions are increasingly embraced and become an integral part of the treatment options. The results of this meta-analysis also suggest that some components of internet interventions for mental disorders, such as recruitment strategy, symptom severity, guidance, and length of the intervention, are worth considering upon implementation. This is the first study to pool cost-effectiveness outcomes in the area of psychiatry, paving the way for other researchers to apply this method to new meta-analyses to further enhance the understanding of the cost-effectiveness of internet interventions compared to alternatives. For example, the current study only considered control conditions, as internet interventions are often used for early detection and underserved populations, for which care as usual and a waiting list are realistic comparators. Future work could compare internet interventions with active comparators such as face-to-face treatment, to clarify the difference in effects and costs between these forms of treatment. Additionally, this study looked broadly at the topic of cost-effectiveness of internet interventions for mental disorders and should be considered a starting point. It might be valuable to investigate a study sample with more homogeneous interventions or designs for more precise estimates of cost-effectiveness. Relatedly, modelling studies were not considered in this meta-analysis as pooling outcomes from such studies with those obtained from RCTs was undesirable. Performing a meta-analysis on modelling studies is feasible, however, and might be an interesting extension. Importantly, no study from a non-Western culture or low-income country was included in the study, while especially low-income countries or people in areas where health care is not paid for by the

government might benefit from internet interventions. Initiating internet interventions in such contexts might be a challenge, but can help to reduce health care costs in the long term. Conducting research on the cost-effectiveness of internet interventions in low-income countries is therefore highly commended.

Conclusion

Pooling outcomes of 37 studies revealed a small benefit of internet interventions for mental disorders compared to control conditions. Perspective of the economic evaluation, targeted mental disorder and WTP for one QALY moderated results. Generalizability to new studies is poor given the large variance of the outcome of interest and heterogeneity between studies. The findings show that cost-effectiveness of internet interventions for mental disorders compared to a do-nothing or care-as-usual approach is likely, but not guaranteed. Continuation of high quality and adequately powered economic evaluations is necessary. This is the first study in the area of psychiatry to pool cost-effectiveness outcomes in an aggregate-data meta-analysis, paving the way for other researchers to use and expand this method.

Chapter 5

The ICECAP-A instrument for capabilities: assessment of construct validity and test–retest reliability in a general Dutch population

Published as: Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., Essers, B. A. B., Van Furth, E. F., Spinhoven, P., & Van den Akker-van Marle, M. E. (2022). The ICECAP-A instrument for capabilities: assessment of construct validity and test–retest reliability in a general Dutch population. *Quality of Life Research, 31*, 687–696. https://doi.org/10.1007/s11136-021-02980-5

Abstract

Purpose: The ICEpop CAPability measure for Adults (ICECAP-A) assesses five capabilities that are important to one's well-being. The instrument might be an important addition to generic health questionnaires when evaluating quality of life extending beyond health. This study aimed to conduct a psychometric assessment of the Dutch translation of the ICECAP-A.
Methods: Construct validity of the instrument was assessed in two ways. First, by measuring correlations with the EQ-5D-5L questionnaire and a measure of self-efficacy and, second, by investigating the ability to distinguish between groups known to differ on the construct the ICECAP-A means to capture. Additionally, test–retest reliability was evaluated.
Results: In total, 1002 participants representative of the general Dutch population completed an online survey. For test–retest reliability, 252 participants completed the same questionnaire 2 weeks later. The ICECAP-A indicated moderate to strong correlations with the EQ-5D-5L and a strong correlation with self-efficacy. Furthermore, it was capable of differentiating known groups. Moreover, results indicated adequate test–retest reliability with an intraclass correlation coefficient of 0.79.
Conclusion: In summary, results suggest adequate test–retest reliability and construct validity and indicate that the ICECAP-A might be of added value, especially when considering areas outside of the traditional health intervention model.

Plain English summary

It is important to be able to precisely measure quality of life, because that helps in assessing how effective a treatment is. The ICEpop CAPability measure for Adults (ICECAP-A) is a questionnaire that was developed to capture one's quality of life in terms of general well-being. This study aimed to further clarify what the ICECAP-A exactly measures and whether it can do so reliably. That would help to decide when this questionnaire should be used. The main finding of the study is that the ICECAP-A questionnaire indeed captures a concept (related to, but different from physical health) best described as well-being. It does so in a valid and reliable way. This suggests that the ICECAP-A questionnaire can be used to measure quality of life. It will be especially useful in contexts outside the area physical health, such as public health, social care, chronic illness, and mental health.

Introduction

Generic health questionnaires are often used to measure benefits of interventions, even in situations where relevant improvements might not be captured in terms of health. As such, they are criticized to employ a narrow view on quality of life, with emphasis on physical aspects of health and current functional abilities rather than resources, coping capabilities, and general well-being (Byford & Sefton, 2003; Carr-Hill, 1989; Coast, 2004; Pietersma et al., 2013). Certain aspects of quality of life that fall beyond physical health might be underestimated, such as living situations, social support systems, psychological resilience, and the capability to cope with illness. Consequently, this can lead to an undervaluation of effect when assessing the benefits of an intervention, especially in the context of social care, mental health (Goranitis et al., 2016; Mitchell et al., 2017), public health, general well-being, chronic illness, and elderly care. The ICEpop CAPability measure for Adults (ICECAP-A) (Al-Janabi et al., 2012) assesses one's quality of life in terms of capabilities and might be better suited than generic health questionnaires in cases that do not fit the traditional health intervention model. Establishing the reliability and validity of the ICECAP-A is vital in order to confidently use this instrument in studies as a complement to generic health questionnaires (i.e., when changes or improvement in outcomes beyond health alone are expected).

Afentou and Kinghorn (2020) have systematically reviewed the literature for studies exploring the psychometric properties of the ICECAP-A. Included studies suggested the ICECAP-A to be positively correlated with concepts such as feelings of happiness and freedom (Al-Janabi et al., 2013) and moderately or strongly related to health-related quality of life instruments (Chen et al., 2018; Engel et al., 2017). Helter et al. (2020) found similar results concerning the psychometric qualities of the ICECAP-A in a more general systematic review on the use of capability instruments in economic evaluations. Overall, the evidence suggests adequate content and construct validity of the ICECAP-A. Its construct seems to be related to quality of life as measured by generic health questionnaires, albeit conceptually different (Afentou & Kinghorn, 2020). Few studies have investigated the test–retest reliability of the ICECAP-A (Al-Janabi et al., 2015; Holst-Kristensen et al., 2020), so more information on this parameter is required. Additionally, the majority of studies assessing the psychometric properties were conducted in the UK (Afentou & Kinghorn, 2020), the results of which do not

necessarily generalize to translations of the instrument and other countries. At the moment, nine translations of the ICECAP-A exist (i.e., Chinese, Danish, Dutch, French, German, Hungarian, Italian, Persian, and Welsh) and an increasing number of studies is available on the psychometric properties of these translations (Baji et al., 2020; Holst-Kristensen et al., 2020; Linton et al., 2020; Shahtaheri et al., 2020; Tang et al., 2018; Xiong et al., 2021). Assessing the psychometric properties of translations of the ICECAP-A in other countries not only makes it more widely available, but strengthens the confidence in the instrument as a whole. To our knowledge there have been no attempts to assess the psychometric properties of the Dutch translation of the ICECAP-A beyond its face validity (Van Hoof et al., 2016). The current aim of the study is to assess the test–retest reliability and improve the understanding of the construct validity of the Dutch translation of the ICECAP-A.

Methods

Design and participants

A cross-sectional design with an additional test–retest measurement for part of the sample was used to assess the psychometric properties of the ICECAP-A. The sample was recruited by a research market agency as part of a larger study aiming to develop ICECAP-A tariffs for the Dutch general population. A sample representative of the Dutch general population, with differences in residential area, educational level, income, and age, was expected to lead to sufficient variations in well-being for this psychometric assessment. An independent medical ethics committee evaluated the study and confirmed it did not fall under the Medical Research Act, waiving the need for ethical approval (METC Leiden- The Hague-Delft, file number N19.119). Hypotheses for the psychometric assessment of the ICECAP-A were registered at AsPredicted (https://aspredicted.org/blind.php?x=sh4dz6) prior to accessing the data, but after data collection. One analysis on convergence and four tests on known-group differences were added later (not preregistered) in order to improve the interpretability of the measurement properties of the ICECAP-A.

Measurements

Demographics

Extracted information on demographics was (1) age in years, (2) current living region or province, (3) gender, (4) highest completed education level with nine categories (ranging from 'no education' to 'university') that were later transformed to lower, middle, and higher education, (5) employment status with eight categories ranging from 'unemployed' to 'retired', (6) marital status, and (7) household composition. Furthermore, seven questions likely related to experienced well-being were assessed, namely (1) general happiness on a 4-point scale, (2) general health on a 5-point scale, (3) chronic illness (yes/no) and (4) whether this illness obstructs daily life in any way (yes/no), (5) the amount of visits to a general practitioner or other doctor, (6) if there were any hospital visits in the last 3 months (yes/no), and (7) if there were any hospital stays in the last 3 months (yes/no).

ICECAP-A

The ICECAP-A (Al-Janabi et al., 2012) measures five capabilities important to one's quality of life: (1) stability - the extent to which someone can feel settled and secure; (2) attachment - the extent to which someone can feel love, friendship, and support; (3) autonomy - the extent to which someone can feel independent; (4) achievement - the extent to which someone can experience achievement and success; (5) enjoyment - the extent to which someone can experience enjoyment and pleasure. Four levels are available for each of the five capabilities, ranging from [1] not being able to experience a capability at all to [4] being able to fully experience a capability. The ICECAP-A attempts to capture the extent to which one experiences the freedom to be or carry out what one wishes. ICECAP-A scores were transformed into capability values using tariffs for the Dutch general population (accepted for publication), ranging from 1 (full capability) to 0 (no capability).

EQ-5D-5L

The EQ-5D-5L (EuroQol Group, 1990) consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with five levels for each dimension (ranging from "no problems" to "extreme problems/unable to"). Using empirical valuations of the Dutch general public (Versteegh, Vermeulen, et al., 2016) the 3125 possible health states can be transformed to a unique utility score, ranging from 1 (perfect health) to −0.446 (worse than death) and anchored at 0 (death). The EQ-5D also contains a visual analogue scale which records subject's self-reported health on a vertical scale ranging from 0 (worst health you can imagine) to 100 (best health you can imagine). For the current study the scale was presented horizontally rather than vertically, to make the question work better on mobile phone.

Self-efficacy

Self-reported efficacy was assessed with three questions on a 4-point scale (1 = often, 2 = sometimes, 3 = rarely, 4 = never) regarding the feeling that one's life is full with possibilities, the feeling to have no control over one's life, and the feeling that one can do the things one wants to do. The second question was recoded to match the direction of the other two questions, so lower scores reflected higher self-reported efficacy. The sum score (ranging from 3 to 12) was used in construct validity analyses. Additionally, for analyses on known-group differences participants who scored '1' or '2' on all three questions were compared to all other participants.

Study procedures

Individuals willing to participate were informed about the study and asked for informed consent. They could continue to the questionnaires only after consent was obtained. Information the researchers received from the marketing bureau was anonymous and could not be traced back to individuals. Additionally, a part of the sample who completed the first questionnaire were asked to fill out the same questionnaire after 2 weeks to determine test–retest reliability of the ICECAP-A. At the start of this second assessment participants were asked

whether they had experienced a change in health since the previous assessment. Procedures for obtaining informed consent and data handling for the second questionnaire were equal to the first.

Statistical analyses

Reliability

The Intraclass Correlation Coefficient (ICC) was used as index of reliability, since it incorporates both degree of agreement and correlation between measurements. The appropriate approximation of the ICC for the test–retest reliability of the ICECAP-A was calculated following the guideline of Koo and Li (2016). Specifically, a two-way mixed-effects model based on single measurement and aiming for absolute agreement was used to calculate the ICC for ICECAP-A capability values between measurement one and two. An ICC of 0.50–0.75, 0.75–0.90, and greater than 0.90 are considered as moderate, good, and excellent reliability respectively (Koo & Li, 2016).

EQ-5D-5L and self-efficacy correlations

Construct validity of the ICECAP-A was evaluated in two ways. First, by investigating correlations of the ICECAP-A with self-efficacy and the EQ-5D-5L. Second, by examining known-group differences. A list of all hypotheses on construct validity can be found in Appendix D.1. Hypothesis 1 (H1) concerned the correlation between ICECAP-A capability values and utility scores of the EQ-5D-5L. While both instruments aim to capture different constructs (i.e., well-being and health), the comparison is relevant to better understand if and when the ICECAP-A can complement generic health measures.

It was expected that the anxiety/depression subscale of the EQ-5D correlated with all subscales of the ICECAP-A (H2–H6), because one of the presumptions of the ICECAPA is that it is specifically suitable for people with mental health complaints (Mitchell et al., 2017). Higher levels of anxiety/depression were expected to relate to lower scores on the ICECAP-A subscales. Five hypotheses were based on earlier findings that the achievement and, especially, autonomy attributes of the ICECAP-A might relate more strongly to physical health than the other three attributes (Keeley et al., 2016). Specifically, we expected that having problems concerning mobility, self-care, and usual activities (EQ-5D) would be reflected in lower autonomy scores on the ICECAP-A (H7–H9). Additionally, we expected that reporting problems concerning usual activities and having pain on the EQ-5D would relate negatively to achievement on the ICECAP-A (H10 and H11). Lastly, as chronic pain (Kawai et al., 2017) and leisure time and activities (Iso-Ahola & Mannell, 2004) are related to life enjoyment we expected that having problems concerning usual activities and having pain (EQ-5D) would make it more difficult for people to experience enjoyment and pleasure (ICECAP-A; H12 and H13). For all hypotheses we expected a significant medium to high correlation ($0.3<r<0.7$) in the direction explained above. The upper boundary to the correlation was set, because we expected the questionnaires and subscales to be related, but also conceptually distinct. Other correlations between the ICECAP-A and EQ-5D subscales were explored, but there were no predetermined expectations.

Lastly, a strong correlation between the ICECAP-A capability values and the self-efficacy sum scores was expected (H14). Spearman rho correlations were used for all hypotheses, since variables were measured at an ordinal level. Multiple testing was accounted for using Holm's method (Holm, 1979).

Known-group differences

Another way of validating the ICECAP-A is to examine its ability to distinguish groups which we know or expect to differ on the construct that the ICECAP-A tries to capture. First, the level of agreement between the two measurements of the ICECAP-A was calculated to give an indication of the stability of repeated scores within participants. Similar to the method used in Gärtner et al. (2015) the standard error of measurement (SEM) was used as an indicator of level of agreement. The SEM constitutes the standard deviation of measurement error and can be derived from the error variance of an analysis of variance for repeated measures, including systematic differences: $SEM = \sqrt{\sigma^2_{time} + \sigma^2_{error}}$. After calculating the SEM of the ICECAP-A capability values differences of known groups were calculated. For a hypothesis to be confirmed the differences need to be both statistically significant and greater than the SEM. Known groups were based on self-reported happiness ratings, the visual analogue scale of the EQ-5D, the presence of a chronic illness, the impeding quality of the illness, visits to a general practitioner, visits to a hospital, hospital stays, self-reported self-efficacy, employment status, marital status, and education (H16–26). Details on the hypotheses can be found in Appendix D.1. Hypotheses 16–25 were tested with the Mann–Whitney U-test and hypothesis 26 with the Kruskal Wallis test, since the ICECAP-A capability values did not follow a normal distribution. Multiple testing was accounted for using Holm's method.

Sample size

The desired sample size for analyses concerning construct validity including known-group differences was 1000, since then even small correlations (e.g., 0.2) can be determined with high precision (e.g., .06) (Cohen, 1988). For test–retest reliability a sample size of 248 was intended. This would yield a power of 0.9, when the acceptable and expected ICC were estimated to be 0.7 and 0.8 relatively, participants were rated twice and 20% of the participants would not qualify for test–retest analyses (Walter et al., 1998).

Results

Participants

Of the 1002 participants who completed the first assessment, 252 also completed the second assessment. Data from the first assessment were used for investigation of construct validity. Mean completion time of the survey was 13.9 min ($SD = 28.0$; range 3.8–618.4). Participants who completed the first assessment within five minutes ($N = 61$) were excluded from analyses, due to concerns with regard to the validity of the results. All participants were invited to complete the second assessment, but the assessment was closed when 250 responses

were gathered. Data from the second assessment were used for test–retest reliability analysis. On average there were 26.7 days ($SD = 2.5$) between the first and second assessment. No time limit was set for the second assessment, since it was very brief. However, participants who indicated to have experienced a change in their health ($N = 44$) were excluded from test–retest analysis, since this analysis assumes conditions for participants have remained the same. Finally, data of 941 and 208 participants were used for construct validity and reliability analyses respectively. Characteristics of all included participants are shown in Table 1. Additionally, a comparison of the sample with the Dutch general population can be found in Appendix D.2.

Table 1. Means and frequencies of participant characteristics

Variable	Category	Construct validity sample (T1; N=941)	Test-retest sample (T2; N=208)
Age		49.4 (17.1)	56.0 (16.1)
Gender	Female	484 (51.4%)	95 (45.7%)
	Male	455 (48.4%)	113 (54.3%)
	Other	2 (0.2%)	0 (0%)
Education	Primary and/or lower education	192 (20.4%)	52 (25.0%)
	Secondary and/or vocational education	395 (42.0%)	76 (36.5%)
	Higher and/or college education	353 (37.5%)	80 (38.5%)
Marital status	Single	186 (19.8%)	32 (15.4%)
	Living together/married/ registered partner	590 (62.7%)	137 (65.9%)
	Relationship	50 (5.3%)	6 (2.9%)
	Divorced	74 (7.9%)	21 (10.1%)
	Widow/widower	33 (3.5%)	9 (4.3%)
	Other	8 (0.9%)	3 (1.4%)
Self-efficacy		5.87 (1.86)	–
ICECAP-A	Capability value	0.88 (0.14)	0.90 (0.13)
EQ-5D-5L	Index scores	0.85 (0.20)	0.86 (0.21)
	Visual analogue scale	76.4 (20.1)	77.3 (19.2)

Note. Values represent mean values with standard deviations in parentheses unless indicated otherwise.

Test-retest reliability

The mean change in ICECAP-A capability value between assessment one and two of the 208 included participants was −.006 ($SD = .084$). For the 44 excluded participants who reported a change in health since the previous assessment the mean change in ICECAP-A capability values was −.015 ($SD = .082$). This indicates that the change in ICECAP-A values for these participants was larger than for the included participants who reported no change in health, but still small. The ICC was 0.79 with a 95% confidence interval (CI) of

0.73—0.84, indicating good test–retest reliability. In comparison, the ICC of the EQ-5D was 0.79 (95% CI 0.74—0.84). Reliability estimates and level of agreement for individual items of the ICECAP-A and EQ-5D are presented in Appendix D.3. The results suggest moderate reliability of individual items of the ICECAP-A.

Construct validity

Correlations with the EQ-5D-5L and self-efficacy

Mean capability values of the ICECAP-A and index scores of the EQ-5D-5L can be found in Table 1 and details concerning individual item frequencies of the questionnaires can be found in Appendix D.3. Fourteen hypotheses were tested to investigate the construct validity of the ICECAP-A. Results on all construct validity hypotheses can be found in Table 2 and the correlation matrix between subscales of the ICECAP-A and EQ-5D-5L can be found in Appendix D.4. Mainly, a substantial Spearman correlation between the ICECAP-A capability values and EQ-5D index scores was found ($r = 0.60$). Additionally, the self-efficacy measure showed a strong Spearman correlation of 0.63 with the ICECAP-A capability values, while its correlation with the EQ-5D-5L index scores was less strong ($r = 0.52$). In total, 12 of 14 (86%) were confirmed.

Table 2. Results on hypotheses for construct validity

Hypo-thesis	ICECAP-A scale	Comparator	Spearman's rho	p-value	Confirmed
H1	Capability value	EQ-5D-5L Index score	0.60	$< .001$	Yes
H2	Stability	Anxiety/depression[a]	0.50	$< .001$	Yes
H3	Attachment	Anxiety/depression[a]	0.44	$< .001$	Yes
H4	Autonomy	Anxiety/depression[a]	0.33	$< .001$	Yes
H5	Achievement	Anxiety/depression[a]	0.38	$< .001$	Yes
H6	Enjoyment	Anxiety/depression[a]	0.49	$< .001$	Yes
H7	Autonomy	Mobility[a]	0.25	$< .001$	No
H8	Autonomy	Self-care[a]	0.27	$< .001$	No
H9	Autonomy	Usual activities[a]	0.44	$< .001$	Yes
H10	Achievement	Usual activities[a]	0.48	$< .001$	Yes
H11	Achievement	Pain/discomfort[a]	0.41	$< .001$	Yes
H12	Enjoyment	Usual activities[a]	0.37	$< .001$	Yes
H13	Enjoyment	Pain/discomfort[a]	0.34	$< .001$	Yes
H14	Capability value	Self-efficacy	0.63	$< .001$	Yes

[a] Subscale of the EQ-5D-5L

Known-group differences

The SEM, based on mean ICECAP-A capability values of the first and second assessment, equalled .0039. This equals 0.39% of the ICECAP-A capability value range, going from 0 to 1. In other words, based on our sample a difference between groups on the ICECAP-A capability value of .0039 or smaller can be attributed to measurement error, while bigger

Table 3. Results on hypotheses for known-group differences

Hypo-thesis	Known group	N	Mean rank score	Median	Range	p-value	Confirmed
H16	Happy	800	515	0.9428	$0.0 - 1.0$	$< .001$	Yes
	Unhappy	141	219	0.7562	$0.3 - 1.0$		
H17	VAS ≥ 65	714	540	0.9448	$0.4 - 1.0$	$< .001$	Yes
	VAS < 65	227	255	0.7879	$0.0 - 1.0$		
H18	No illness	562	564	0.9495	$0.4 - 1.0$	$< .001$	Yes
	Illness present	379	334	0.8546	$0.0 - 1.0$		
H19[a]	Non-obstructing illness	51	255	0.9226	$0.5 - 1.0$	$< .001$	Yes
	Obstructing illness	328	180	0.8312	$0.0 - 1.0$		
H20	No hospital visit	588	511	0.9375	$0.2 - 1.0$	$< .001$	Yes
	Hospital visit	353	405	0.9149	$0.0 - 1.0$		
H21	No hospital stay	860	477	0.9305	$0.0 - 1.0$	$= .017$	Yes
	Hospital stay	81	402	0.9149	$0.4 - 1.0$		
H22	No GP visit	383	549	0.9475	$0.2 - 1.0$	$< .001$	Yes
	GP visit	558	417	0.9149	$0.0 - 1.0$		
H23	High self-efficacy	415	601	0.9565	$0.5 - 1.0$	$< .001$	Yes
	Low self-efficacy	526	368	0.8790	$0.0 - 1.0$		
H24	Employed	811	501	0.9375	$0.0 - 1.0$	$< .001$	Yes
	Unemployed/ occupational disability	130	283	0.8144	$0.2 - 1.0$		
H25	Relationship	640	504	0.9375	$0.3 - 1.0$	$< .001$	Yes
	No relationship	301	401	0.9070	$0.0 - 1.0$		
H26[b]	Higher education	353	NA	0.9339	$0.2 - 1.0$	$= .021$	No
	Medium education	395		0.9339	$0.3 - 1.0$		
	Lower education	192		0.9149	$0.0 - 1.0$		

GP general practitioner; *VAS* visual analogue scale of the EQ-5D-5L

[a] This question was only applicable to 379 participants who indicated to have a chronic illness

[b] One subject is missing from this analysis since the response to this question was not interpretable

differences are likely due to actual differences between groups. Results on all known-group hypotheses can be found in Table 3. In summary, 10 of 11 (91%) of hypotheses were confirmed. For education, a significant difference was found between groups, but only lower and higher education had a capability value difference that was both significant ($p = .005$) and larger than the SEM, contradicting expectations. The other known-group differences were significant and larger than the SEM, confirming the predetermined hypotheses. Known-group hypotheses were repeated with the EQ-5D-5L index scores to get a better understanding of the difference between the EQ-5D-5L and the ICECAP-A. Results on these analyses can be found in Appendix D.5. Both questionnaires performed similarly in distinguishing known groups. When looking at the size of the median difference between tested known groups in relation to the SEM the EQ-5D-5L might distinguish groups based on hospital visits and hospital stays more clearly than the ICECAP-A, while the ICECAP-A might be especially good in distinguishing groups based on happiness, overall health (based on EQ-5D-5L VAS scores), self-efficacy, employment, and relationship status.

Discussion

The aim of this study was to assess the psychometric properties of the ICECAP-A in a large sample representative of the general Dutch population. The instrument showed good test–retest reliability with an ICC of 0.79. Good construct validity was found based on correlations with the EQ-5D-5L and a measure of self-efficacy, with 12 of 14 hypotheses (86%) being confirmed. Similarly, the ICECAP-A showed adequate construct validity by being able to differentiate between known groups, with 10 of 11 hypotheses (91%) being confirmed.

In general, correlations between the ICECAP-A and EQ-5D-5L were moderate to strong. This result suggests that while there is considerable overlap between the two instruments, there may be a difference in the underlying measured constructs. Interestingly, the correlation between the autonomy subscale of the ICECAP-A and the EQ-5D subscales self-care and mobility was poor (smaller than 0.3, though still significant). This is surprising given that difficulties with moving and taking care of oneself imply that help from others is needed. It might be that such difficulties can be overcome without help from others, through the use of (walking) aids or extra effort, or that aspects of autonomy not related to physical capabilities, such as being able to make choices, explain the variance on the autonomy item better. Another explanation is that a ceiling effect on the EQ-5D dampened the correlation. Indeed, 70% and 91% of the participants reported the highest level of mobility (i.e., 'no problems with walking') and self-care (i.e., 'no problems with washing and getting dressed'), respectively. For the autonomy subscale of the ICECAP-A considerably less participants (48%) reported the highest level (i.e., 'able to be completely independent'). Overall 33% of the participants reported the maximum score on the EQ-5D, whereas 14% did so for the ICECAP-A. This suggests that the ICECAPA, compared to the EQ-5D, might have more room to detect subtle changes in quality of life. This heightened sensitivity has been established in other populations (Goranitis et al., 2016; Mitchell et al., 2017).

Contrary to our hypothesis, the difference in capability value did not exceed the SEM while also being significant for all three educational groups. Only the comparison between

higher and lower educational groups fulfilled both criteria. The hypothesis was based on earlier research indicating that the EQ-5D could discriminate similar groups (Janssen et al., 2012), but an additional analysis suggested that the EQ-5D, compared to the ICECAP-A, performed roughly equal in discriminating the three educational groups in the current sample. Regarding other known-group differences, the EQ-5D-5L seemed to distinguish groups more clearly than the ICECAP-A when groups were based on hospital visits and hospital stays. This seems further evidence that the EQ-5D-5L puts more emphasis on health, while the ICECAP-A has a broader focus. Indeed, the ICECAP-A distinguished groups more clearly when groups were based on concepts related to general well-being, such as happiness, relationship status, and self-efficacy. These results are in line with earlier research suggesting that the ICECAP-A correlated positively with feelings of happiness and freedom (Al-Janabi et al., 2013). Moreover, the self-efficacy measure correlated strongly with the ICECAP-A capability value, indicating they measured overlapping concepts. The substantial correlation should not be surprising, since self-efficacy is defined as an individual's belief about their own capabilities and mastery over their life (Bandura, 1988) which seems very similar to the construct of the ICECAP-A as described by the developers (Al-Janabi et al., 2012).

Previous research and implications

Regarding test–retest reliability, similar results were established in a previous studies. A slightly higher ICC of 0.86 for the ICECAP-A capability values was found in a sample from the Danish population (Holst-Kristensen et al., 2020) and an ICC of 0.72 was found in a general UK sample (Al-Janabi et al., 2015). In this UK study, reliability of the ICECAP-A was found to be lower than the EQ-5D, which might be explained in part by the inherent property of capabilities being harder to objectify than health. Indeed, the current study also showed a lower test–retest reliability of individual items of the ICECAP-A compared to those of the EQ-5D. However, no difference between the ICC estimates of the ICECAP-A capability values and EQ-5D index scores was found.

The same research team also found comparable results regarding validity (Al-Janabi et al., 2013). In a sample of 418 participants representative of the general UK population 97 hypotheses were formed regarding construct validity of which 67 (69%) were confirmed. It must be noted that multiple comparisons were not accounted for, which likely increased the amount of significant findings. Nevertheless, the authors stated that while their research does not indicate definitive validity of the ICECAP-A, it does show potential in capturing intervention benefits because of its ability to identify relevant differences between groups. This statement is solidified in other studies. For example, in a substance dependence sample Goranitis et al. (2016) found that the ICECAP-A has stronger correlations than the EQ-5D with concepts that are often important objectives of interventions, such as social support, functioning, and well-being. Additionally, compared to the EQ-5D the ICECAP-A was found to be more sensitive to change, which has been reproduced in a sample with depression (Mitchell et al., 2017), and advocates its use in samples suffering from chronic or mental disorders. However, this does not mean that capability instruments like the ICECAP-A should replace health questionnaires like the EQ-5D. Combining previous findings with that of the current study suggests that the ICECAP-A will perform especially well in contexts outside of the traditional health intervention model, while generic health questionnaires will

do better when health is the outcome of interest. Indeed, previous studies (Engel et al., 2017; Keeley et al., 2016) and the NICE social care guidelines (National Institute for Health and Care Excellence, 2016) suggest that the two instruments assess different constructs and can effectively complement each other. The Dutch guidelines for conducting economic evaluations in healthcare also specify that the ICECAP should be added when interventions aim to improve not only health gain, but well-being in terms of living situation, autonomy, and social interaction as well (Zorginstituut Nederland, 2015).

Strengths, limitations, and future directions

A strength of this psychometric evaluation was that the study was preregistered to ensure reliable hypotheses testing. Secondly, appropriate statistical choices were made such as using a suitable ICC, correcting for multiple testing, and examining both the significance and size of correlations and differences. Thirdly, a large sample representative of the general Dutch population was used. Quotations based on age, gender, and income were used during recruitment, resulting in a heterogeneous sample regarding health, well-being, happiness, and education level, and a good starting point for assessing psychometric properties. Future studies exploring the responsiveness of the ICECAP-A should consider more specific populations.

Admittedly, some limitations can be indicated. First, the ICECAP-A was administered online only so results do not necessarily generalize to a paper–pencil version of the questionnaire. However, there are no reasons to expect a difference between the two methods and earlier work confirms this for the EQ-5D-5L (Lundy et al., 2020). Second, for construct validity the ICECAP-A was compared to the EQ-5D-5L and a measure of self-efficacy. Including other quality of life, health or capability instruments, and assessment of discriminative validity might have led to an enhanced understanding of the psychometric properties of the ICECAP-A. Nevertheless, the current analyses add to the understanding of the ICECAP-A construct and its added value to health-related quality of life measures. Third, regarding test–retest reliability, there was on average 26.7 days between assessment one and two which may have introduced recall bias. Lastly, changes in well-being at the second assessment were assessed by asking participants whether they had experienced a change in health since the previous assessment rather than also informing on changes in well-being. While there was a larger decline in ICECAP-A capability values in the group who reported a change in health since the first assessment, the change was still small, questioning the appropriateness of this check of changes in well-being.

Conclusion

Adequate psychometric properties of the ICECAP-A are vital to be able to reliably use the instrument. The present study adds to the established literature on the psychometric properties of the ICECAP-A by showing good test–retest reliability and construct validity in a large Dutch sample. The instrument demonstrates both overlap and differences with the EQ-5D-5L, indicating that the ICECAP-A might measure a distinct concept, closely related to well-being and self-efficacy, that is influenced by health status. Consequently, the

ICECAP-A can complement other generic health questionnaires when attempting to capture the benefits of interventions outside the traditional health intervention model.

Chapter 6

The ICEpop Capability Measure for Adults instrument for capabilities: development of a tariff for the Dutch general population

Published as: Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G., Van Til, J. A., Essers, B. A., Van Furth, E. F., & Van den Akker-Van, M. E. (2022). The ICEpop Capability Measure for Adults Instrument for Capabilities: Development of a Tariff for the Dutch General Population. *Value in Health, 25*, 125-132. https://doi.org/10.1016/j.jval.2021.07.011

Abstract

Objectives: The ICEpop Capability Measure for Adults (ICECAP-A) assesses 5 capabilities (stability, attachment, autonomy, achievement, and enjoyment) that are important to one's quality of life and might be an important addition to generic health questionnaires currently used in economic evaluations. This study aimed to develop a Dutch tariff of the Dutch translation of the ICECAP-A.
Methods: The methods used are similar to those used in the development of the UK tariff. A profile case best–worst scaling task was presented to 1002 participants from the general Dutch population. A scale-adjusted latent class analysis was performed to test for preferences of ICECAP-A capabilities and scale heterogeneity.
Results: A 3-preference class 2-scale class model with worst choice as scale predictor was considered optimal and was used to calculate the resulting tariff. Results indicated that the capabilities stability, attachment, and enjoyment were considered more important aspects of quality of life than autonomy and achievement. Additionally, improving capabilities from low to moderate levels had a larger effect on quality of life than improving capabilities that were already at a higher level.
Conclusion: The ICECAP-A tariffs found in this study could be used in economic evaluations of healthcare interventions in The Netherlands.

Introduction

Efficient allocation of resources is becoming increasingly important when it comes to making decisions in healthcare and health policy. Cost-utility analysis is a central tool for judging the efficiency of interventions and can support decisions on healthcare funding. Generally, quality-adjusted life-years (QALYs) are the central outcome measure in cost-utility analyses. To assess quality of life, generic utility measures are often used, such as the EQ-5D (EuroQol Group, 1990) or the Short-Form 6 Dimensions (Brazier et al., 1998). Nevertheless, there is critique on the use of generic health questionnaires for economic evaluations, mainly that not all relevant domains of quality of life are captured by these instruments (Byford & Sefton, 2003; Carr-Hill, 1989; Coast, 2004). Indeed, Pietersma et al. (2013) analyzed several generic utility measures and found that they capture only a selective amount of domains of quality of life and use an almost exclusive focus on people's current functional abilities with little emphasis on coping capabilities and resources. Consequently, relevant benefits of interventions outside the area of physical health might be underestimated in current economic evaluations.

Accordingly, considering a different, broader approach not limited to health-related quality of life might be more appropriate for determining treatment outcomes, especially for patients with a psychiatric disorder (Mitchell et al., 2017) or chronic illness. One such approach is based on capabilities (Sen, 1992, 1993). Capabilities indicate the extent to which someone is able to do what one wishes to do. The ICEpop Capability Measure for Adults (ICECAP-A) (Al-Janabi et al., 2012) is an instrument that measures well-being based on capabilities and may be an appropriate addition to the established EQ-5D. The instrument is receiving increased international recognition (Flynn et al., 2015) and may be used for economic evaluations of treatments aimed at improving not only physical health but well-being in general. Indeed, regarding its construct, existing research suggests that the ICECAP-A correlates positively with concepts such as feelings of happiness and freedom (Al-Janabi et al., 2013) and that it can capture information beyond health-related quality of life (Afentou & Kinghorn, 2020; Engel et al., 2017; Keeley et al., 2016). Economic evaluations that have already been conducted with the ICECAP-A suggest that using capabilities might lead to different decisions on resource allocation (Al-Janabi et al., 2013).

To be able to use the ICECAP-A in economic evaluations tariffs are needed to translate answers of patients on the ICECAP-A to a capability value between "0" and "1," where "0" represents "not at all able to do what one wishes" and "1" represents "fully able to do what one wishes." These anchoring values are different to utility values where "0" represents "health as bad as death" and "1" represents "perfect health." Tariffs of the ICECAP-A of a certain population indicate how important the various capabilities are according to that population and they might differ between populations, cultures, and countries. A tariff already exists for the general population of the United Kingdom (Flynn et al., 2015), but to be able to reliably use the ICECAP-A in other countries, tariffs for those countries need to be developed. In The Netherlands, using the ICECAP in economic evaluations is recommended when benefits regarding well-being are expected (Zorginstituut Nederland, 2015), but no Dutch tariff is available. This study aimed to develop an ICECAP-A tariff for the Dutch general population.

Methods

Design, Participants, and Procedure

Methods used to establish the Dutch tariff of the ICECAP-A are similar to those used for the development of the UK tariff by Flynn et al. (2015). Participants were approached by a market research agency (Kantar Group). A sample of 1002 participants was recruited that was representative of the Dutch general population based on age, gender, region, and income. Because questionnaires were completed online with less possibility for guidance throughout the assessment compared with the interviews in the study by Flynn et al. (2015) and a larger study size was recommended by the UK research group, a sample size of 1000 was assumed to be adequate for establishing the Dutch tariff. Additionally, Yang et al. (2015) showed that in discrete choice experiments a sample size of 1000 provides sufficient power for study designs that were similar to that of the current study (type 2 best–worst scaling, conditional logit latent class model) in terms of estimator properties. Participants were first informed about the study and could only continue to the online questionnaire if they consented with participating. They were paid a small sum of money to complete the questionnaire. Only fully completed assessments were saved and no information on the amount or content of partially completed questionnaires was stored. Information the researchers received from the marketing bureau was anonymous and could not be traced back to individuals. An independent medical ethics committee evaluated the study and confirmed it did not fall within the Medical Research Act, waiving the need for ethical approval (Medisch Ethische Toetsingscommissie Leiden-The Hague-Delft, file number N19.119).

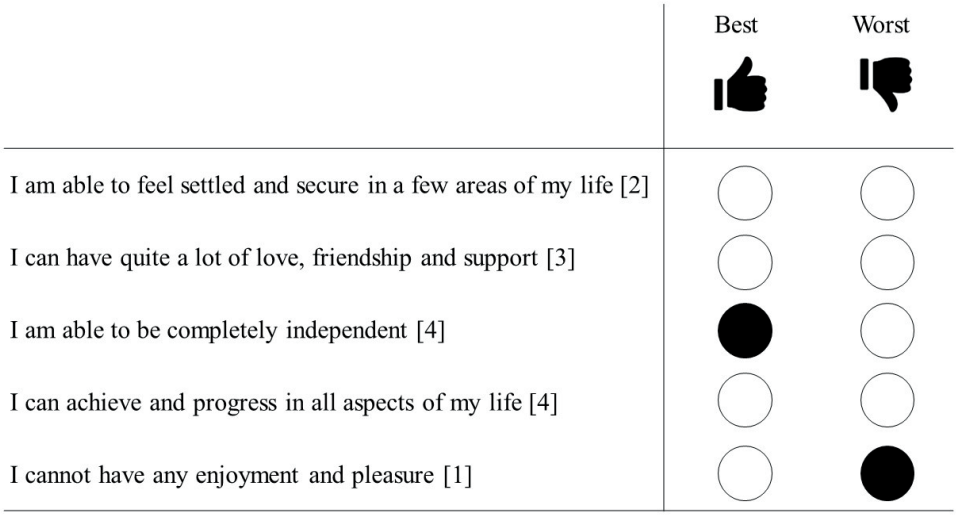
Measurements

Best-worst scaling task

The ICECAP-A comprises 1024 (4 levels for each capability) possible states. Using the orthogonal main-effect plan (OMEP) design created by Flynn et al. (2015), 16 profiles, each containing 1 possible ICECAP-A state, were presented to participants. Half of the participants were presented with the 16 profiles from the OMEP design and the other half with its 16 foldover profiles (e.g., capabilities presented at level 4, 3, 2, or 1 in the original OMEP design were presented at level 1, 2, 3, and 4, respectively, in this foldover). The OMEP design and its foldover can be found in Appendix E.1. For each of the 16 profiles, participants had to indicate which of the capabilities they valued as best and which as worst. This is known as a (profile case) best–worst scaling task (Potoglou et al., 2011). An example of a profile can be seen in Figure 1. A pilot questionnaire was completed in an in-person interview by a convenience sample of 10 people of different ages and educational level to confirm the task would be understood by participants.

In the final questionnaire, participants were first asked to complete questions on demographics and their health and the ICECAP-A (Flynn et al., 2010). Details on these questionnaires can be found in Appendix E.2. Here, the levels of capabilities (shown behind every statement) were presented. Participants rated the experienced difficulty of completing the ICECAP-A on a 4-point scale (ranging from 1 "very easy" to 4 "very difficult"). Then, based on experiences from the pilot, an explanation of the best–worst scaling task was given

Figure 1. Example of a completed best–worst profile.



Note. Sixteen such profiles were completed in Dutch by participants. The number in straight brackets [#] indicates the level of the corresponding statement, ranging from [1], the lowest level, to [4], the highest level. In the example, the participant evaluated statement 3 "completely independent" to be the best (i.e., adds the most to a valuable life) and statement 5 "cannot have any enjoyment and pleasure" to be the worst (i.e., obstructs having a valuable life the most).

with an example of one completed profile. The explanation and best–worst scaling task can be found in Appendix E.3.

Statistical Analyses

Best-worst pairs table

Firstly, a table was constructed with all possible best–worst pairs. In other words, a count was made of how often, for example, stability at level 1 was chosen as best, whereas attachment at level 1 was chosen as worst, which resembled 1 of the 320 possible best–worst pairs. The margins of the table provided an initial understanding of the perceived importance to quality of life of the 20 capability levels. Moreover, the table allowed inspection of the frequencies of unlikely choices (e.g., attributes presented at level 4 chosen as worst or attributes presented at level 1 chosen as best), providing insight into the quality of the data.

Best-minus-worst scores

Second, best-minus-worst scores for participants showed individual preferences for capability levels and were used to estimate choice consistency. Within the OMEP design (and its

foldover), each capability level was presented 4 times. The best-minus-worst score for 1 capability level, then, equaled the times that a participant picked that capability level as best minus the times it was picked as worst. This resulted in 20 best-minus-worst scores ranging from $+4$ (0 times picked as best and 4 times picked as worst) to -4 (4 times picked as best and 0 times picked as worst). Next, for each individual, the sum of squares for each capability was used to calculate the empirical scale parameter (ESP), which gave an indication of the consistency with which a participant made choices. An ESP (ranging from 0 to 8) of approximately 4 was considered normal for a participant who understood the task and made consistent choices (Flynn et al., 2015). Participants with a suspicious answering pattern on the best–worst scaling task, identified by differing more than 2 standard deviations (SDs) from the average ESP, were excluded from analyses concerning the tariff development. Table 1 depicts a set of best-minus-worst scores of a participant to illustrate the calculations.

Table 1. Best-minus-worst scores for one of the participants.

Capability	Level	Best-minus-worst score	Normalized ($*1/4$) and squared	Sum of squares
Stability	1	-3	0.56	1.38
	2	0	0	
	3	2	0.25	
	4	3	0.56	
Attachment	1	-2	0.25	0.88
	2	-1	0.06	
	3	0	0	
	4	3	0.56	
Autonomy	1	-1	0.06	0.38
	2	-2	0.25	
	3	1	0.06	
	4	0	0	
Achievement	1	-1	0.06	0.19
	2	0	0	
	3	1	0.06	
	4	1	0.06	
Enjoyment	1	-3	0.56	1.44
	2	-2	0.25	
	3	1	0.06	
	4	3	0.56	
ESP				4.25

ESP empirical scale parameter

Scale-adjusted latent class analysis

Latent Gold 5.1 software was used for scale-adjusted latent class (SALC) analysis. These analyses can distinguish individuals with different preferences (i.e., preference heterogeneity) by adding preference classes and also individuals with similar preferences but with different

choice consistency (i.e., scale heterogeneity) by adding scale classes (Magidson & Vermunt, 2007). Although SALC models are not the only option to model both preference and scale heterogeneity, they are widely used and unique in estimating separate classes with differing preferences (Groothuis-Oudshoorn et al., 2018). As new preference classes are added to the model, the software uses the data to predict the probability for an individual to fall within a certain class. Each class has its own parameters (comparable with regression coefficients) for each of the 20 capability levels of the ICECAP-A, where parameters further away from 0 signify greater importance (i.e., are more often chosen as best or worst than other capability levels). Effects coding was used with level 4 of enjoyment as reference level. Adding more classes to a model will often improve the fit, but a balance between fit and interpretability is warranted. Nevertheless, there are no clear guidelines for choosing one model over another. Therefore, we chose to follow a pragmatic approach by, on one hand, minimizing the Bayesian information criterion (BIC) and, on the other hand, looking for a solution with classes that were clearly separable. Apart from adding classes, it is possible to add scale classes to separately target scale heterogeneity (Vass et al., 2018). For people in the same class but in a different scale class, parameters of capability levels showed a similar pattern, but were scaled. The scaling factor was smaller than 1 if they were less consistent or larger than 1 if they were more consistent in making best–worst choices. Additionally, to account for possible heteroscedasticity (i.e., allow a different scale factor) between best and worst choices, a dummy variable indicating a worst choice was added as scale predictor to the estimated models. Finally, multiple starting seeds were used when estimating the SALC model to verify the stability of the solution.

In the final model, the relative attribute importance within each class gave an indication of the preferences of participants in that class. Attribute importance was calculated for the five attributes in all classes by dividing the parameter range of one ICECAP-A attribute (i.e., the difference between level 1 and level 4 parameters of an attribute) by the sum of five attribute parameter ranges.

ICECAP-A tariff

After identifying the preferred model, the parameters of each class and scale class were weighted by the size of the class (i.e., the probability that a participant falls into that particular class) by calculating the product of the raw parameters and the group probability. Finally, adding the weighted parameters for every capability level across groups resulted in 20 parameters that, when linearly transformed to range from 0 (i.e., level [1] for all capabilities) to 1 (i.e., level [4] for all capabilities), constituted the final tariff.

Results

Participants

In total, 1002 participants completed the online questionnaire. The distribution of the ESP can be found in Appendix E.4. The ESP differed 2 SDs from the mean (4.04, $SD = 1.18$) for 69 participants (40 below and 29 above the mean). Visual inspection confirmed that these participants had suspicious answering patterns (e.g., always choosing stability as best

and enjoyment as worst, regardless of the level on which they were presented) suggesting they did not understand the task or did not take it seriously. These participants were excluded, leaving 933 participants for analyses. Excluding these participants did not influence representativeness of the sample (see Appendix E.5) or the balance between randomization to version 1 and 2 of the best–worst scaling task (50.1% vs 49.9%) and had a small effect on quality of the data (see Appendix E.6). The questionnaire took on average 14.2 minutes ($SD = 28.9$, range $3.8 - 618.4$) to complete. One participant for whom completion time was 5692 minutes was not included in this calculation. There were no missing data. Table 2 presents participant characteristics. The sample was highly representative of the general Dutch population in terms of age, gender, region, and income (see Appendix E.5). Most participants found the ICECAP-A very easy or easy to complete (93.9%).

Table 2. Frequencies (%) and means (standard deviations) of participant characteristics.

Variable	Category	Sample mean ($N = 933$)
Age		48.9 (17.1)
Gender	Female (%)	479 (51.3)
	Male (%)	453 (48.6)
	Other (%)	1 (0.1)
ICECAP-A	Capability value*	0.88 (0.13)
ICECAP-A difficulty	Very easy (%)	469 (50.5)
	Easy (%)	407 (43.6)
	Hard (%)	55 (5.9)
	Very hard (%)	2 (0.2)
EQ-5D-5L	Index scores*	0.86 (0.20)
ESP		4.07 (0.95)

Note. Values represent mean values with standard deviations in parentheses unless indicated otherwise.
ESP Empirical scale parameter.
*Values reflect scores based on the Dutch population tariff.

Best-Worst Pairs Table

The number of times each of the 320 best–worst pairs was chosen across all participants is presented in Table 3. The last column indicates how often a capability at a certain level is chosen as best, whereas the last row indicates how often a capability at a certain level is chosen as worst. For example, the capability attachment presented at level 4 ("I can have a lot of love, friendship and support") was chosen 1772 times (11.9% of best choices) as best and 229 times (1.5% of worst choices) as worst across all profiles that participants completed. The table suggests that high levels of stability, attachment, and, to a lesser extent, autonomy and enjoyment were often chosen as best, whereas high levels of achievement were infrequently chosen as best (9.7%, 11.9%, 7.2%, and 8.7%, respectively, vs 3.5%). For worst choices, preferences appeared less explicit, with low levels of stability, attachment, autonomy, achievement, and enjoyment all frequently chosen as worst (10.5%, 9.8%, 9.3%, 8.3%, and 10.1%, respectively).

Table 3. Best–worst pairs frequencies.

Best		Worst																				Total	% (best choices)
		Stability				Attachment				Autonomy				Achievement				Enjoyment					
	Level	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4		
Stability	1	×	×	×	×	15	14	9	12	14	11	9	8	16	15	13	14	22	8	10	8	198	1.33
	2	×	×	×	×	64	19	18	25	56	22	26	18	33	28	37	29	55	27	27	13	497	3.33
	3	×	×	×	×	207	47	19	17	134	73	34	56	171	114	43	55	193	78	24	30	1295	8.67
	4	×	×	×	×	179	61	27	15	159	151	46	33	134	66	104	82	184	151	25	26	1443	9.67
Attachment	1	5	5	9	10	×	×	×	×	9	8	13	12	28	22	23	17	14	7	10	5	197	1.32
	2	80	29	18	7	×	×	×	×	39	28	22	30	55	60	42	42	107	24	14	14	611	4.09
	3	165	76	18	33	×	×	×	×	252	98	51	48	191	158	81	77	149	78	18	21	1514	10.14
	4	233	143	30	24	×	×	×	×	169	121	34	47	170	174	113	91	236	143	24	20	1772	11.87
Autonomy	1	7	9	10	9	13	5	7	10	×	×	×	×	9	12	13	10	9	6	11	5	145	0.97
	2	72	13	12	17	49	24	15	16	×	×	×	×	35	24	27	25	97	22	18	19	485	3.25
	3	199	87	19	39	148	40	18	15	×	×	×	×	83	60	61	36	121	101	23	27	1077	7.21
	4	141	56	25	21	101	119	9	13	×	×	×	×	79	109	54	47	133	111	31	27	1076	7.21
Achievement	1	9	6	9	3	17	11	18	12	6	5	8	11	×	×	×	×	4	6	5	4	134	0.90
	2	60	14	7	19	31	18	10	19	18	14	10	15	×	×	×	×	44	10	11	8	308	2.06
	3	87	45	7	17	132	38	20	18	49	18	19	16	×	×	×	×	105	32	12	8	623	4.17
	4	65	60	11	14	64	54	12	14	59	51	14	17	×	×	×	×	40	29	10	12	526	3.52
Enjoyment	1	14	5	5	6	8	5	8	6	7	10	5	4	11	4	5	7	×	×	×	×	110	0.74
	2	56	23	15	18	59	14	7	10	50	22	22	23	34	16	20	33	×	×	×	×	422	2.83
	3	194	60	23	16	219	58	5	14	211	86	24	37	91	70	41	43	×	×	×	×	1192	7.98
	4	180	134	29	33	162	60	29	13	162	82	36	42	91	133	63	54	×	×	×	×	1303	8.73
Total		1567	765	247	286	1468	587	231	229	1394	800	373	417	1231	1065	740	662	1513	833	273	247	29856	100.0
% (worst choices)		10.5	5.1	1.7	1.9	9.8	3.9	1.6	1.5	9.3	5.4	2.5	2.8	8.3	7.1	5.0	4.4	10.1	5.6	1.8	1.7	100.0	

Note. Based on $N = 933$. Row margins indicate best choice frequencies and column margins indicate worst choice frequencies.

SALC Estimates

A 3-preference class 2-scale class model with worst choice as a scale predictor was considered optimal ($df = 871$, $BIC = 68992$, $R^2 = 0.25$). A 3-preference class was chosen because a third class added a substantial group with interpretable differences compared with a 2-preference class model ($df = 894$, $BIC = 71166$, $R^2 = 0.19$). Adding a fourth class resulted in one relatively small group that did not provide clear discrimination between already existing preference classes ($df = 854$, $BIC = 69224$, $R^2 = 0.25$). Two scale classes were added because they improved the fit of the model considerably. Adding a third scale class reduced both the fit and the interpretability of the model. All attribute parameters for participants in the second scale class were estimated to be 0.29 times those of participants in the first scale class, with most participants (58.1%) predicted to be in the first scale class. Finally, adding worst choice as a scale predictor increased the fit of the model and seemed relevant to control for the questionnaire design (where participants could pick the best and worst choice in whatever order they preferred). Indeed, the scaling factor for worst choices compared with best choices was 0.68 ($p < .001$). This suggests that participants switched the order of making best and worst choices throughout the best–worst scaling task, strengthening the choice to correct for questionnaire design by adding worst choice as a predictor in the model. Relatedly, a strong linear relation between the amount of best choices and the inverse of worst choices across each of the 20 capability levels was found ($r = 0.97$, $R^2 = 0.95$), indicating that best and worst data were proportional and can likely be pooled for analyses. A summary of the results on all estimated models can be found in Appendix E.7.

A table with attribute importance, based on the parameters from Table 4, can be found in Appendix E.8. Participants in preference class 1, containing 40.2% of the sample, showed little variation in attribute importance with stability, attachment, autonomy, achievement, and enjoyment accounting for 0.23, 0.20, 0.21, 0.17, and 0.20 of the space, respectively. Participants in class 2, containing 30.3% of the sample, were characterized by a very low preference for achievement ($.02$) with high preferences for the other four capabilities. Class 3 contained 29.5% of the participants and was distinguished by a high preference for attachment (0.30) and enjoyment (0.27) while indicating low importance of autonomy (0.14) and especially achievement ($.09$). For the total sample, the attribute importance for stability, attachment, autonomy, achievement, and enjoyment weighted by class size was 0.22, 0.24, 0.19, 0.13, and 0.22, respectively.

Table 4. Final model parameters and Dutch general population ICECAP-A tariffs.

	Class 1 sClass 1	Class 1 sClass 2	Class 2 sClass 1	Class 2 sClass 2	Class 3 sClass 1	Class 3 sClass 2	Final Dutch tariff
Class probability	0.2337	0.1686	0.1761	0.1270	0.1712	0.1234	
	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	
Stability [1]	25.84 (.33)	21.71 (.04)	20.68 (.14)	20.20 (.20)	23.86 (.13)	21.13 (.08)	−0.0073

Stability [2]	21.02	20.30	0.52	0.15	20.63	20.19	0.1061
	(.12)	(.03)	(.04)	(.12)	(.10)	(.04)	
Stability [3]	3.17	0.93	1.37	0.40	1.97	0.58	0.2007
	(.18)	(.03)	(.07)	(.13)	(.10)	(.05)	
Stability [4]	4.22	1.23	1.34	0.39	2.09	0.61	0.2163
	(.20)	(.03)	(.09)	(.14)	(.10)	(.05)	
Attachment [1]	25.11	21.50	20.71	20.21	24.45	21.30	−0.0035
	(.31)	(.04)	(.13)	(.20)	(.13)	(.09)	
Attachment [2]	20.83	20.24	0.76	0.22	0.43	0.13	0.1223
	(.11)	(.03)	(.04)	(.11)	(.10)	(.03)	
Attachment [3]	2.71	0.79	1.56	0.46	3.49	1.02	0.2118
	(.18)	(.03)	(.07)	(.13)	(.11)	(.06)	
Attachment [4]	3.80	1.11	1.59	0.47	4.17	1.22	0.2344
	(.19)	(.03)	(.09)	(.15)	(.12)	(.08)	
Autonomy [1]	25.30	21.55	21.21	20.35	23.07	20.90	0.0027
	(.32)	(.04)	(.13)	(.18)	(.12)	(.08)	
Autonomy [2]	20.90	20.26	0.33	0.10	20.78	20.23	0.1043
	(.11)	(.03)	(.04)	(.11)	(.10)	(.03)	
Autonomy [3]	2.69	0.79	0.97	0.29	0.87	0.26	0.1784
	(.16)	(.03)	(.06)	(.12)	(.11)	(.04)	
Autonomy [4]	3.88	1.14	0.69	0.20	0.86	0.25	0.1920
	(.19)	(.04)	(.09)	(.17)	(.13)	(.05)	
Achievement [1]	24.41	21.29	21.80	20.53	22.56	20.75	0.0143
	(.29)	(.04)	(.11)	(.17)	(.13)	(.06)	
Achievement [2]	20.90	20.26	21.69	20.49	20.94	20.28	0.0813
	(.12)	(.04)	(.04)	(.12)	(.14)	(.04)	
Achievement [3]	1.76	0.52	21.54	20.45	0.08	0.02	0.1308
	(.14)	(.04)	(.05)	(.12)	(.13)	(.03)	
Achievement [4]	2.90	0.85	21.63	20.48	0.02	0.00	0.1451
	(.19)	(.04)	(.07)	(.14)	(.13)	(.04)	
Enjoyment [1]	25.25	21.53	21.10	20.32	24.14	21.21	−0.0063
	(.31)	(.04)	(.13)	(.19)	(.12)	(.08)	
Enjoyment [2]	21.64	20.48	0.09	0.03	0.07	0.02	0.1001
	(.14)	(.03)	(.05)	(.12)	(.12)	(.04)	
Enjoyment [3]	2.57	0.75	0.59	0.17	2.87	0.84	0.1932
	(.17)	(.04)	(.07)	(.12)	(.13)	(.06)	
Enjoyment [4]*	3.48	1.02	0.54	0.16	3.53	1.03	0.2122
	(.19)	(.04)	(.08)	(.14)	(.13)	(.07)	

Note. Scale factor sClass 2 compared with sClass 1 = 0.2925.
Coef Coefficient, *sClass* Scale class
*Used as reference level

ICECAP-A Tariff for the General Dutch Population

Table 4 shows the coefficients for the different preference classes and scale classes, together with the tariff. The capability value can be deduced from the tariff by adding the values for the corresponding score. For example, a change in an ICECAP-A score of [12211] to [44323]

would result in a change in capability value of 0.6762: from 0.2274 $(-0.0073 + 0.1223 + 0.1043 + 0.0143 - 0.0063)$ to 0.9036 $(0.2163 + 0.2344 + 0.1784 + 0.0813 + 0.1932)$. The capability value was scaled to range from 0 [11111] to 1 [44444].

In the chosen model, the capability attachment on level 4 was valued as most desirable (parameter $= 2.28$, tariff $= 0.2344$) and capability stability on level 1 as least desirable (parameter $= -2.60$, tariff $= -0.0073$) to one's quality of life. The largest increase in capability equals 0.1258 and is obtained when going from attachment level 1 ("cannot have any love, friendship, and support") to level 2 ("can have a little love, friendship, and support"). The average difference between capability levels was 0.0667. The largest relative importance was ascribed to attachment, accounting for 22.3% of the possible improvement, whereas achievement received the lowest preference, accounting for 13.1% of the possible improvement. In general, the capabilities stability, attachment, and enjoyment seem to be somewhat more important to quality of life than autonomy and achievement. In addition, improvements within a capability from a low level to a higher level (e.g., going from level 1 to 2) yielded larger increases in capability value than improving attributes that were already moderate to high (e.g., going from level 3 to 4).

Explorative analyses were conducted after developing the tariff to investigate what aspects of quality of life are important for different people. Details on these explorative analyses can be found in Appendix E.9.

Discussion

This study aimed to develop a tariff for the ICECAP-A based on a large representative sample from the general Dutch population ($N = 933$). The tariff shows that the five capabilities described in the ICECAP-A all contribute to quality of life. The capabilities stability, attachment, and enjoyment were somewhat more important than autonomy, and achievement contributed the least to quality of life. Going from one level to the next within an attribute does not have a linear effect on the tariff. Indeed, improving capabilities from low to moderate levels rather than from moderate to high is more valuable according to the current sample. Consequently, prioritizing to help people with low capabilities might result in larger well-being gains for society as a whole. This relates to the concept of "sufficient capability," an approach with the aim to maximize the number of people above a level of sufficient capability (Goranitis et al., 2017; Mitchell et al., 2015).

Most study findings are similar to those reported for the UK tariff (Flynn et al., 2015). It is to be expected that Dutch and UK populations have comparable preferences. Nevertheless, it is interesting to note that the Dutch sample seems to value high levels of enjoyment more and high levels of achievement less compared with the UK sample. This difference in preferences was also apparent in the European Values Study (European Values Study, 2017), where 95.6% of Dutch respondents indicated that leisure time is important in their lives compared with 91.9% of their UK counterparts. More strikingly, 81.0% of UK respondents indicated that the feeling to achieve something is an important aspect of a job, whereas this was only the case for 62.4% of the Dutch respondents. Consequently, interventions that increase the ability to enjoy life might have a slightly greater impact on quality of life in The Netherlands than the United Kingdom. Capturing these differences between countries

in tariffs is important because they might ultimately influence funding decisions (Kiadaliri et al., 2015).

Strengths and Limitations

The SALC model used to find clusters of participants with similar answer patterns is a flexible model that enables the modeling of both preference and scale heterogeneity, resulting in a parsimonious model. The BIC was used to determine the final model. Nevertheless, this measure tends to overstate the number of preference classes (Groothuis-Oudshoorn et al., 2018), so the final model was also based on interpretability and face validity, inevitably introducing subjective judgment. Another choice was to use case 2 (profile) best–worst scaling to establish participant preferences on the ICECAP-A. It must be noted that although best–worst scaling tasks might be more statistically efficient than discrete choice experiments, estimates of preferences seem to be similar across methods (Whitty & Gonçalves, 2018) and evidence on the burden on participants is mixed (Flynn et al., 2007; Himmler et al., 2021; Mühlbacher et al., 2016). A strength of the study was the recruitment of a large sample to develop the tariff.

Several limitations were also present. First, people with lower education were somewhat underrepresented because the assessment was online and education was not included in the quotations. Additionally, the sample was slightly under representative of the 75- to 99-year age group. Possibly, this is related to a difficulty of finding participants in this age group with access to the internet. These differences between the sample and the Dutch population might have influenced the tariff slightly. Second, a pilot was conducted to identify problems and to assess the difficulty of the best–worst task, which led to significant improvements in explanations in the questionnaire. Nevertheless, the final questionnaire was completed online with no guidance making it impossible to check how participants interpreted the questions. At least 69 participants did not understand the task or take it seriously and were excluded from analyses, but it is realistic to assume that more participants struggled with the questionnaire. Indeed, the margins of best–worst pairs table reveal that in the remaining sample 12% of worst choices were a capability presented at level 4 and 5% of best choices were a capability presented at level 1. This is strange considering all profiles presented to participants had balanced capability levels with some capabilities presented at a high level and others at a low level. Nevertheless, because the conducted analyses could account for scale heterogeneity and the sample was large with the majority seeming to understand the task, it is expected that the current results still reflect preferences on quality of life of the Dutch general population accurately.

Use in Economic Analyses

To be able to compare (economic) benefits across interventions, it is necessary to consider both the effectiveness (i.e., quality of life) and life extension (i.e., quantity of life). Conceptually, it is difficult to interpret the capability value derived from tariffs of the ICECAP-A in the context of health economics and cost-utility analyses and in comparison with QALYs (Coast, Smith, et al., 2008; Cookson, 2005). The capability value is not a QALY because the lowest value is not anchored to "death," but to "no capability." Nevertheless, death

is accounted for in the sense that death is associated with no capability even though the reverse is not necessarily true (e.g., consider a state in which capabilities are nonexistent or a state of unconsciousness) (Coast, Flynn, et al., 2008). Consequently, capability values have a meaningful anchor (i.e., no capability) and can be adjusted for time, by estimating gains in years lived with full capability (Flynn et al., 2015). Therefore, they can be used in economic evaluations in a similar way as QALYs. Although applied similarly, the ICECAP-A measures a related but distinct concept compared with generic health questionnaires (Afentou & Kinghorn, 2020). This suggests that the ICECAP-A is not a substitute, but rather a complement to generic health questionnaires (Engel et al., 2017; Keeley et al., 2016), as is also advocated by the National Institute for Health and Care Excellence social care guidelines (National Institute for Health and Care Excellence, 2016). Accordingly, the instrument seems to be especially suitable and valuable in contexts outside the traditional healthcare model, such as general well-being, social care, mental health (Goranitis et al., 2016; Mitchell et al., 2017), public health, and chronic illness. Indeed, the Dutch guidelines for conducting economic evaluations in healthcare recommend the use of the ICECAP when considering interventions aimed at improving general well-being (Zorginstituut Nederland, 2015).

Conclusion

This study developed a tariff for the ICECAP-A based on a large Dutch general population. This makes the ICECAP-A ready for use in economic evaluations in The Netherlands. The instrument is expected to be a valuable addition to other generic health questionnaires, especially when evaluating interventions outside the traditional health intervention model.

Chapter 7

Cost-effectiveness of three internet-based interventions for eating disorders: results from a randomized controlled trial

Published as: Rohrbach, P. J., Dingemans, A. E., Van Furth, E. F., Spinhoven, P., Van Ginkel, J. R., Bauer, S., & Van den Akker-Van Marle, M. E. (2022). Cost-effectiveness of three internet-based interventions for eating disorders: results from a randomized controlled trial. *International Journal of Eating Disorders, 55*, 1143-1155. https://doi.org/10.1002/eat.23763

Abstract

Objective: The primary aim was assessing the cost-effectiveness of an internet-based self-help program, expert-patient support, and the combination of both compared to a care-as-usual condition.
Method: An economic evaluation from a societal perspective was conducted alongside a randomized controlled trial. Participants aged 16 or older with at least mild eating disorder symptoms were randomly assigned to four conditions: (1) Featback, an online unguided self-help program, (2) chat or e-mail support from a recovered expert patient, (3) Featback with expert-patient support and (4) care-as-usual. After a baseline assessment and intervention period of 8 weeks, five online assessments were conducted over 12 months of follow-up. The main result constituted cost-utility acceptability curves with QALYs and societal costs over the entire study duration.
Results: No significant differences between the conditions were found regarding QALYs, health care costs and societal costs. Non-significant differences in QALYs were in favor of the Featback conditions and the lowest societal costs per participant were observed in the Featback only condition (€16,741) while the highest costs were seen in the care-as-usual condition (€28,479). The Featback only condition had the highest probability of being efficient compared to the alternatives for all acceptable willingness-to-pay values.
Conclusion: Featback, an internet-based unguided self-help intervention, was likely to be efficient compared to Featback with guidance from an expert patient, guidance alone and a care-as-usual condition. Results suggest that scalable interventions such as Featback may reduce health care costs and help individuals with eating disorders that are currently not reached by other forms of treatment.

Public significance statement

Internet-based interventions for eating disorders might reach individuals in society who currently do not receive appropriate treatment at low costs. Featback, an online automated self-help program for eating disorders, was found to improve quality of life slightly while reducing costs for society, compared to a do-nothing approach. Consequently, implementing internet-based interventions such as Featback likely benefits both individuals suffering from an eating disorder and society as a whole.

Introduction

Eating disorders are burdensome in terms of disability, quality of life and mortality (Arcelus et al., 2011; Smink et al., 2013) and also from an economic perspective (Erskine et al., 2016; Van Hoeken & Hoek, 2020). There exists a large treatment gap for eating disorders, meaning that many individuals with an eating disorder do not get help specifically for their eating disorder, despite having substantial symptoms (Austin et al., 2021). Nevertheless, they generally make more use of health care services compared to people without an eating disorder (Hart et al., 2011; Van Son et al., 2012; Weissman & Rosselli, 2017), which is reflected in higher health care costs (Ágh et al., 2016; Samnaliev et al., 2015). Eating disorder related costs may become larger still when also considering costs outside health care, such as productivity losses and caregiver costs (Deloitte Access Economics, 2020). These substantial costs warrant well-advised resource allocation decisions. In fact, investing in evidence-based treatment for eating disorders might ultimately result in cost savings (Bode et al., 2017). Apart from such policy changes, helping individuals with an eating disorder while reducing costs for society requires continued effort from researchers and clinicians to make treatments more effective, accessible and less expensive.

Cost-effectiveness research, where outcomes and costs of two different courses of action are compared, is necessary to distinguish interventions that are more efficient than others. Internet interventions, often coined as cost-effective alternative to other treatment options, have frequently been confirmed in their effectiveness (Linardon et al., 2020; Loucas et al., 2014; Melioli et al., 2016; Pittock et al., 2018), but cost-effectiveness research in scarce. Across mental disorders, evidence from systematic reviews cautiously suggests internet-based interventions might indeed be cost-effective, at least compared to do-nothing approaches (Ahern et al., 2018; Donker et al., 2015; Paganini et al., 2018). A few studies investigated the cost-effectiveness of internet interventions compared to face-to-face eating disorder treatment (Crow et al., 2009; König et al., 2018; Watson et al., 2018) and found internet interventions to be slightly less effective in reducing eating disorder symptoms, but also less costly. Consequently, such interventions might be especially efficient as a first step in a stepped-care treatment model, as they have the potential to reach individuals that currently do not get appropriate care for their eating disorder (Aardoom, Dingemans, & Van Furth, 2016). When researching these first step internet-based interventions for eating disorders, care as usual may be used as a reference, since it represents the, often inappropriate, care individuals with eating disorders in society receive. Unfortunately, there is a paucity of cost-effectiveness research comparing online interventions for eating disorders with care

as usual. A simulation study on US college students with eating disorders indicated that a stepped-care treatment model with online guided self-help was less costly and resulted in fewer individuals in need of additional treatment than usual care (Kass et al., 2017). Recently, Akers et al. (2021) showed an online version of the cognitive-dissonance based intervention 'the Body Project' to have health benefits compared to enhanced usual care, while health utilization was similar. Additionally, Aardoom, Dingemans, van Ginkel, et al. (2016) found that Featback, an online automatic monitoring and feedback system for people with an eating disorder, was cost-effective compared to a care-as-usual condition, regardless of whether the intervention was complemented with chat or e-mail support by a psychologist. Taken together, the limited evidence available suggests that online interventions for eating disorders may be cost-effective compared to care as usual, which is especially interesting considering that such interventions are scalable and easily accessible and can reach people at an early stage of eating disorder development. Recently, a second randomized controlled trial to replicate and extend the results on the effectiveness and cost-effectiveness of Featback compared to care as usual was conducted. In the first RCT (Aardoom, Dingemans, Spinhoven, et al., 2016), support by a psychologist did not add to the effectiveness of Featback. Possibly, support by expert patients (i.e., recovered individuals) is more fitting and effective for those reluctant to seek help (Rohrbach et al., 2019).

Aims

The primary aim of this study was to investigate the cost-effectiveness of the three conditions, 1) the fully automated internet intervention Featback, 2) chat or email support from expert patients and 3) the combination of both interventions compared to 4) a care-as-usual condition from a societal perspective. The three active online interventions were expected to be more efficient than care as usual.

Method

Design and randomization

This economic evaluation was part of a randomized controlled trial, pre-registered at the Dutch Trial Register (NL7065) and approved by an independent medical ethics committee (METC-LDD; NL64553.058.18). Detailed information on the interventions and methods can be found in the study protocol (Rohrbach et al., 2019). Results on the clinical effectiveness will be reported elsewhere. A two-by-two factorial design with the internet-based interventions Featback and expert-patient support was used, resulting in four conditions: (1) Featback, (2) Featback with expert-patient support, (3) expert-patient support and (4) care-as-usual condition. All conditions had a duration of eight weeks. Assessments on quality of life and costs were all online and completed by participants at post intervention and 3, 6, 9 and 12 month follow up. Participants were randomized and distributed across conditions in blocks of 40. For randomization, a computer-generated random numbers list was made by an independent researcher, concealing it from the principal investigator before and during the trial. The economic analysis maintains a societal perspective, meaning that both health

care costs and non-healthcare costs were included. Data concerning costs and utility covered a period of 14 months (i.e., 8 weeks intervention or waiting plus 12 months follow up).

Participants

Participants were recruited mainly via Proud2Bme, a Dutch online community for people with eating-related problems or eating disorders, from October 2018 to October 2019. After expressing interest to participate, they received a screening questionnaire. Eligible participants were 16 years or older, had internet access and reported at least mild eating disorder symptoms. Specifically, they scored 52 or higher on the Weight Concerns Scale (Killen et al., 1993) or reported a body mass index lower than or equal to 18.5, or one or more weekly binge eating episodes or compensatory behaviors in the past four weeks on the Short Evaluation of Eating Disorders (Bauer et al., 2005). Participants with severe eating disorder symptoms were advised to seek professional help, but were not excluded as they too may benefit from the offered interventions.

Interventions

Participants in all conditions were free to undergo any other type of intervention or treatment, representing individuals in society with varying levels of treatment. Consequently, the waiting list control condition can be seen as care as usual for individuals with eating disorder symptoms in (Dutch speaking) society.

Featback

Participants could make weekly use of an automated monitoring and feedback system for eight weeks. Based on the answers of a short monitoring questionnaire, participants received a supportive feedback message with a summary of self-reported eating problems, psychoeducation, and guidance on how to counter eating disorder related symptoms. Current level of impairments as well as improvements or deteriorations in eating disorder related symptoms compared to the previous week were captured in the messages. Additionally, participants could access the Featback website with psycho-educative material on eating disorders at their own convenience.

Expert-patient support

Five expert patients (sometimes referred to as peers or mentors) were recruited, who had a lived experience of an eating disorder and were fully recovered. They received a protocol and were trained on how to use their own experience to help others overcome their eating disorder via chat and e-mail. Monthly supervision from an experienced expert patient and clinical psychologist during the trial was included. Participants allocated to the conditions with expert-patient support were assigned to one of the expert patients for eight weeks and could schedule a 20-minute chat or e-mail session every week. Chat sessions closed automatically after 20 minutes. For e-mail sessions, participants sent an e-mail to their expert patient before the scheduled time slot and the expert patient responded during the appointment.

Featback with expert-patient support

Participants in this condition were able to make use of both Featback and weekly 20-minute chat or e-mail support from an expert patient.

Care-as-usual condition

Participants in this condition were placed on a waiting list for 14 months. After the waiting period, participants were offered eight weeks of Featback with weekly expert-patient support.

Measures

Demographics

Assessed baseline variables were age, gender, nationality, education level, eating disorder treatment history, marital status, weight, height, eating disorder duration, internet usage, eating disorder symptoms assessed with the Eating Disorder Examination Questionnaire global scores (Fairburn & Beglin, 2008).

Quality of life

The primary outcome measure for the economic evaluation was quality-of-life adjusted life years (QALYs) as assessed with the EQ-5D-5L (EuroQol Group, 1990), which demonstrates adequate psychometric properties (Feng et al., 2021). The Dutch tariff (Versteegh, Vermeulen, et al., 2016) was used to translate EQ-5D-5L scores to utility values. Subsequently, QALYs were calculated over the 14 month follow-up period using the area-under-curve method.

Because generic health questionnaires like the EQ-5D-5L might be limited in their extent to detect changes in wellbeing for interventions aimed at mental health (Pietersma et al., 2013) the economic evaluation was also conducted using the ICECAP-A (Al-Janabi et al., 2012). Psychometric properties of the ICECAP-A have been found to be adequate (Afentou & Kinghorn, 2020; Rohrbach, Dingemans, Essers, et al., 2022). A capability value anchored at 0 (no capability) and 1 (full capability) was calculated for each participant using the ICECAP-A Dutch tariffs (Rohrbach et al., 2021) over the 14 month study period. Details on the used quality-of-life instruments and accompanying transformations can be found in Appendix F.1.

Costs

Health care costs included intervention costs and use of health care services. Intervention costs for Featback included five minutes of technical support by a researcher (including setting up an account, redirecting participants to professional help in the case of severe symptom deterioration and responding to technical problems) multiplied by their hourly rate (€31.50). For expert-patient support, costs were calculated by multiplying their hourly rate (€22.31) with the time spent on support sessions (i.e., estimated at 30 minutes for each session, including preparation and administration). Additionally, supervision costs were calculated by

dividing the total time spent on supervision (i.e., 14 one-hour sessions attended by six expert patients with €22.31 hourly wage, one researcher with €31.50 hourly wage and one clinical psychologist with €106.17 hourly wage) by the amount of participants in the two conditions with the possibility of expert-patient support. All wages were determined based on the real wages during the conduct of the study.

Health care costs were measured with the TiC-P midi (Timman et al., 2015) at each assessment (i.e., over an 8-week period at post intervention and 3 months at all other follow-up assessments). The midi version was chosen over the full version as it reduced the time burden for participants while maintaining a reliable estimate of health service use (Timman et al., 2015). Finally, visits to the general practitioner, dietician, psychologist based in mental health institutions, the private section or hospitals, medical specialist, the emergency department, daycare in mental health institutions, and hospitalizations either in the hospital or a mental health institution were included as health care costs. After inspecting the data for errors and possible double counts, the amount of visits to each health care provider was multiplied with their cost prices as indicated by the Dutch guidelines for cost research in health care (Hakkaart-van Roijen et al., 2015; Kanters et al., 2017). All assessed health care services with their reference price are presented in Table 1.

Non-health care costs were measured with the Productivity Costs Questionnaire (PCQ) (Bouwmans et al., 2015), including costs related to absence from work (absenteeism), reduced productivity at work because of health problems (presenteeism) and reduced productivity of unpaid work such as domestic chores because of health problems. Absenteeism costs were calculated by multiplying the recalled hours of missed work over the last 4 weeks extrapolated to 8 weeks (at post intervention) or 3 months (at follow-up measurements) by the average gross hourly wage of female working individuals in the Netherlands (Hakkaart-van Roijen et al., 2015). In cases of longer absence through illness the friction cost method was applied, meaning that no costs were incurred after being absent for 12 weeks, because initial production levels were expected to have been restored by that time. Presenteeism costs were calculated by multiplying the recalled hours with reduced productivity because of health problems over the last 4 weeks extrapolated to 8 weeks or 3 months by the average gross hourly wage of female working individuals in the Netherlands (Hakkaart-van Roijen et al., 2015). Lastly, costs related to reduced productivity of unpaid work was calculated by multiplying the recalled hours in which others had to perform domestic chores instead of the participant in the last 4 weeks extrapolated to 8 weeks or 3 months by the average gross hourly wage of a domestic worker (Hakkaart-van Roijen et al., 2015). Gross hourly wages are presented in Appendix 2.

All costs were indexed to the year 2021 using the Dutch consumer price index (OECD, 2021b). No discounting was applied to QALYs and costs, given that the time horizon was slightly more than 1 year.

Missing data

Baseline values of the EQ-5D-5L and ICECAP-A were not available. As these variables appear stable over a relatively short period (i.e., 8 weeks) of time, they were estimated to be equal to those at post intervention for the main analyses. This assumption was tested using sensitivity analyses.

Table 1. Price references

Category	Reference prize	CPI index 2014-2021	CPI index 2019-2021	Final cost price (2021)
Intervention costs[a]				
Featback (5min researcher coordination per participant; hourly wage of €30.72	€2.56		1.025	€2.62
Expert-patient support session (30min per session; hourly wage of €22.31)	€11.16		1.025	€11.44
Supervision costs per participant			€21.38	
Direct health care costs				
General practitioner	€33.00	1.095		€36.15
Dietician	€33.00	1.095		€36.15
Psychologist, psychotherapist or psychiatrist - mental health care	€98.00	1.095		€107.35
Psychologist, psychotherapist or psychiatrist - independent	€94.44	1.095		€103.45
Psychologist, psychotherapist or psychiatrist - hospital	€91.00	1.095		€99.68
Medical specialist	€91.00	1.095		€99.68
Emergency department	€259.00	1.095		€283.70
Day treatment - mental health care				€183.05
Hospitalization - mental health care	€302.36	1.095		€331.20
Hospitalization - hospital	€476.00	1.095		€521.40
Indirect costs				
Average gross hourly female wage	€31.60	1.095		€34.61
Average gross hourly domestic worker wage	€14.00	1.095		€15.34

CPI=Cost Price Index.
[a] Wages of the research coordinator, expert patient and clinical psychologist (supervision) were based on the real wages during the conduct of the study.
Note. Dutch CPI indexes for 2021, 2019 and 2014 were €108.88, €106.20 and €99.40 respectively.

According to the intention-to-treat approach, all participants who completed baseline were included throughout the analyses. Missing data were multiply imputed (Rubin, 1986) using the software program R version 3.5.1. Details on the multiple imputation procedure can be found in Appendix F.2.

Statistical analyses

Costs, both health care and societal, and effects in terms of QALYs (EQ-5D-5L) and capabilities (ICECAP-A) over the 14 month period were compared between the four conditions using analyses of variance (ANOVA) pooled across imputations (Rubin, 1987; Van Ginkel & Kroonenberg, 2014). Multiple testing was corrected for using Holm's method (Holm, 1979). Cost-utility analyses were conducted with QALYs and societal costs over the 14 month follow-up period. Specifically, QALYs and costs were averaged over the 100 imputed datasets. Subsequently, a bootstrap procedure simulating 1000 samples drawn from the average imputation sample was conducted in Microsoft Excel to estimate the uncertainty regarding mean costs and QALYs. Mean costs and QALYs per study condition were used to calculate the incremental net benefit (INB) for each condition. To calculate the INB, first, society's willingness to pay (WTP) for one extra year lived in perfect health (i.e., 1 QALY) was multiplied with the QALY gain in a condition, which expresses the effect in monetary terms. Subtracting the costs for this condition resulted in its INB. The 1000 INBs for each condition were used to calculate the probability of a condition to be cost-effective compared to the other conditions for a range of WTP values. In the Netherlands the willingness to pay is assumed to vary between €20,000 per QALY for interventions in the context of 'low disease burden' to €80,000 per QALY in the context of severe diseases (Zwaap et al., 2015). To accommodate all relevant WTP values, the current study explored values ranging from €0 to €100,000. The results were presented in cost-utility acceptability curves for the four conditions separately.

Four sensitivity analyses were conducted to investigate the robustness of the results. Specifically, using the average imputation sample, cost-utility analyses were repeated with (1) capability values based on ICECAP-A scores resulting in cost-capability acceptability curves, (2) QALYs based on utility scores obtained from the visual analogue scale of the EQ-5D-5L (raw scores divided by 100) and (3) direct health care costs only instead of societal costs. Lastly, because baseline scores of the EQ-5D-5L were unavailable, a sensitivity analysis (4) was performed where baseline scores of the EQ-5D-5L were estimated using the 4-item Patient Health Questionnaire (PHQ-4). Equipercentile mapping was used to translate baseline PHQ-4 scores into EQ-5D-5L scores. These were then used to calculate adjusted QALYs for the cost-utility acceptability curves. Details on the mapping procedure can be found in Appendix F.3.

Results

Participants

In total, 355 participants completed informed consent and the baseline assessment and were included in the analyses. Retention of participants at baseline (T0), post intervention (8

weeks; T1) and 3, 6, 9 and 12 month follow-up (T2-T5) was 355 (100%), 280 (78.9%), 252 (71.0%), 244 (68.7%), 233 (65.6%) and 242 (68.2%) respectively. Study drop-out rates did not differ between conditions at post intervention, $\chi^2(3) = 3.99$, $p = .26$, or 12 month follow-up, $\chi^2(3) = 4.90$, $p = .18$. No differences in stopping with the intervention between the three active interventions were found, $\chi^2(2) = 1.24$, $p = .54$. Baseline characteristics of participants are presented in Table 2.

Table 2. Baseline characteristics of participants

Characteristics	Featback ($N = 88$)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)
Gender					
Female (%)	89 (98.9)	82 (93.2)	84 (96.6)	88 (97.8)	343 (96.7)
Male (%)	1 (1.1)	5 (5.7)	1 (1.1)	2 (2.2)	9 (2.5)
Other (%)	0 (0.0)	1 (1.1)	2 (2.3)	0 (0.0)	3 (0.8)
Nationality					
Dutch (%)	80 (88.9)	78 (88.6)	80 (92.0)	81 (90.0)	319 (89.9)
Belgian (%)	9 (10.0)	9 (10.2)	6 (6.9)	8 (8.9)	32 (9.0)
Other (%)	1 (1.1)	1 (1.1)	1 (1.1)	1 (1.1)	4 (1.1)
Education					
Low (%)	12 (13.3)	5 (5.6)	12 (13.7)	18 (20.5)	47 (13.3)
Middle (%)	31 (34.4)	33 (37.5)	34 (39.0)	35 (39.3)	133 (37.6)
High (%)	47 (52.2)	50 (56.8)	41 (47.1)	36 (40.4)	174 (49.2)
Treatment history for ED					
Yes (%)	54 (54.0)	46 (52.3)	53 (60.9)	49 (54.4)	202 (56.9)
No (%)	36 (36.0)	42 (47.7)	34 (39.1)	41 (45.6)	153 (43.1)
Self-reported diagnosis status					
Officially diagnosed with ED	52 (59.1)	60 (66.7)	52 (59.8)	58 (64.4)	222 (62.5)
No diagnosis, but assumed to have ED	24 (27.3)	22 (24.4)	23 (26.4)	22 (24.4)	91 (25.6)
Eating problems, but likely no ED diagnosis	12 (13.6)	8 (8.9)	12 (13.7)	10 (11.1)	42 (11.8)
Marital status					
Married/living together (%)	22 (24.4)	20 (22.7)	26 (29.9)	30 (33.3)	98 (27.6)
Living alone (%)	66 (73.3)	68 (77.3)	58 (66.7)	58 (64.4)	250 (70.4)
Divorced (%)	1 (1.1)	0 (0.0)	3 (3.4)	2 (2.2)	6 (1.6)
Widow (%)	1 (1.1)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.2)
Age [Years]	28.0 (1.7)	28.3 (10.4)	26.8 (9.4)	28.1 (12.4)	27.8 (10.8)

Weight [kg]	64.0 (21.0)	62.2 (18.3)	63.6 (22.0)	64.7 (23.4)	63.6 (21.2)
Height [cm]	169.9 (7.2)	168.5 (6.9)	169.7 (7.1)	169.5 (6.9)	169.4 (7.0)
Years with ED	10.1 (9.1)	10.3 (8.8)	8.6 (8.2)	11.4 (12.0)	10.1 (9.7)
Internet usage [hours per day]	4.2 (2.6)	3.7 (2.2)	3.9 (2.3)	3.4 (2.8)	3.8 (2.5)
Eating disorder symptoms (EDE-Q)	3.9 (1.1)	4.1 (1.1)	4.3 (1.0)	4.3 (1.0)	4.1 (1.0)

Note. Data are presented as means (standard deviation) unless indicated otherwise.
ED = Eating Disorder; EDE-Q = Eating Disorder Examination Questionnaire.

Quality of life

EQ-5D-5L utility values and ICECAP-A index scores for all measurement points, as well as QALYs and capability values over the total study duration (12 months + 8 weeks) are presented in Table 3. Average QALYs were highest in the Featback with expert-patient support condition and lowest for the care-as-usual condition. However, no significant differences in QALYs between the four conditions were found. Similarly, no differences in improvements on capabilities as derived from the ICECAP-A between the four conditions were found.

Costs

Intervention costs, health care costs and non-health care costs are presented in Table 4. Intervention costs were significantly higher in conditions with expert-patient support. Lowest health care costs were found in the Featback only condition, while highest costs were found in the care-as-usual condition. The relatively low health care costs in the Featback only condition could mostly be attributed to fewer participants being hospitalized in that condition. Average societal costs per participant over the study duration were again lowest in the Featback only condition and highest in the care-as-usual condition. Although the omnibus test was significant, after a Holm correction for multiple testing, pooled ANOVA tests revealed no significant difference between the four conditions for health care costs and societal costs.

.

Table 3. Means (standard errors) of utilities, QALYs and capabilities

Category	Featback ($N = 88$)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)	Pooled F-statistic
EQ-5D-5L utilities						
Post intervention	0.68	0.68	0.61	0.58	0.64	$F(3, 333) =$
(T1; 8 weeks)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	3.01, $p = .03$
3-month	0.62	0.68	0.60	0.58	0.62	$F(3, 326) =$
follow-up (T2)	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	1.88, $p = .13$
6-month	0.69	0.69	0.62	0.60	0.65	$F(3, 321) =$
follow-up (T3)	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	1.90, $p = .13$
9-month	0.61	0.65	0.63	0.62	0.63	$F(3, 311) =$
follow-up (T4)	(0.04)	(0.04)	(0.04)	(0.04)	(0.02)	0.19, $p = .91$
12-month	0.66	0.71	0.66	0.64	0.67	$F(3, 317) =$
follow-up (T5)	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	0.61, $p = .61$
EQ-5D-5L Visual Analogue Scale utilities						
Post intervention	0.60	0.59	0.55	0.55	0.57	$F(3, 331) =$
(T1; 8 weeks)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	1.05, $p = .37$
3-month	0.55	0.56	0.55	0.57	0.56	$F(3, 333) =$
follow-up (T2)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	0.22, $p = .88$
6-month	0.61	0.57	0.54	0.58	0.57	$F(3, 325) =$
follow-up (T3)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	1.77, $p = .15$
9-month	0.57	0.59	0.56	0.59	0.58	$F(3, 318) =$
follow-up (T4)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	0.43, $p = .73$
12-month	0.60	0.61	0.57	0.59	0.59	$F(3, 326) =$
follow-up (T5)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	0.52, $p = .67$
ICECAP-A capability values						
Post intervention	0.69	0.68	0.63	0.65	0.66	$F(3, 329) =$
(T1; 8 weeks)	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	1.40, $p = .24$
3-month	0.68	0.69	0.62	0.66	0.66	$F(3, 324) =$
follow-up (T2)	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	1.48, $p = .22$
6-month	0.70	0.69	0.64	0.65	0.67	$F(3, 316) =$
follow-up (T3)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	1.01, $p = .39$
9-month	0.67	0.68	0.65	0.69	0.67	$F(3, 305) =$
follow-up (T4)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	0.44, $p = .72$
12-month	0.72	0.72	0.64	0.72	0.70	$F(3, 313) =$
follow-up (T5)	(0.02)	(0.03)	(0.03)	(0.03)	(0.01)	2.09, $p = .10$
Total QALYs	0.75	0.78	0.71	0.69	0.74	$F(3, 337) =$
EQ-5D-5L[a]	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	1.87, $p = .14$
Total QALYs	0.67	0.64	0.69	0.65	0.66	$F(3, 339) =$
EQ-5D Visual	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	0.74, $p = .53$
Analogue Scale[a]						
Total capability	0.79	0.79	0.73	0.77	0.77	$F(3, 328) =$
values ICECAP-A[a]	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	1.31, $p = .27$

SE=Standard error.
[a] Calculated over the entire 14-month study duration.

Table 4. Means (standard error) of costs per study condition over the course of 14 months in 2021 euros with the percentage of participants that incurred the costs

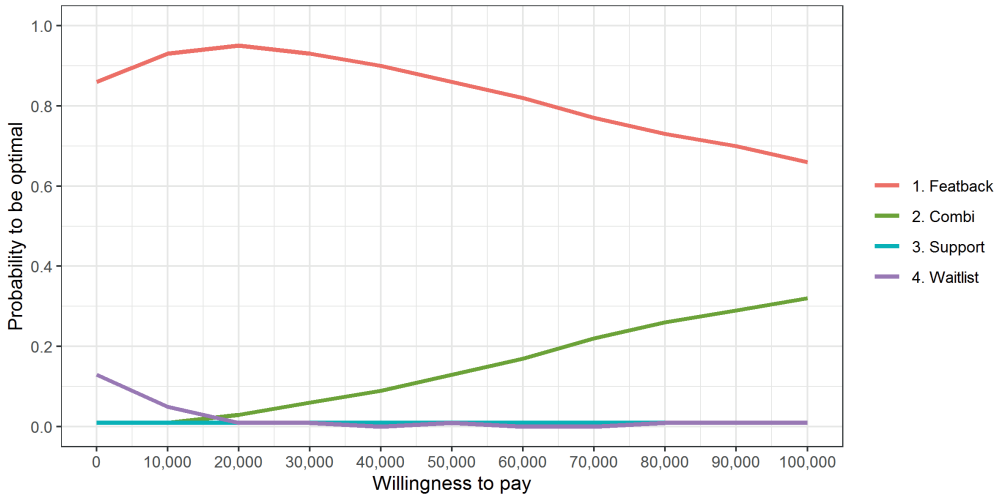
Category	Mean costs per participant (SE) [% of participants incurring costs]					
	Featback ($N = 88$)	Featback + Support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)	Pooled F-statistic
Total intervention costs	3 (0)	65 (4)	72 (4)	0 (0)	35 (2)	$F(3, 351) = 226.45$, $p < .001$
Health care costs						
General practitioner	288 (33) [93%]	260 (39) [92%]	215 (28) [87%]	320 (36) [95%]	271 (17) [92%]	$F(3, 335) = 1.69$, $p = .17$
Dietician	122 (25) [52%]	217 (37) [61%]	133 (31) [53%]	202 (38) [58%]	169 (17) [56%]	$F(3, 334) = 2.12$, $p = .10$
Psych[a] – mental health care	1483 (298) [55%]	1745 (369) [58%]	2723 (448) [70%]	1810 (312) [66%]	1936 (184) [62%]	$F(3, 336) = 2.23$, $p = .08$
Psych[a] - independent	1005 (186) [57%]	869 (153) [59%]	1195 (222) [64%]	1052 (188) [63%]	1029 (95) [61%]	$F(3, 335) = 0.52$, $p = .67$
Psych[a] - hospital	617 (181) [40%]	241 (75) [32%]	714 (188) [43%]	549 (138) [43%]	528 (76) [39%]	$F(3, 340) = 1.83$, $p = .14$
Medical specialist	220 (50) [46%]	225 (47) [56%]	204 (43) [46%]	297 (53) [57%]	237 (24) [52%]	$F(3, 332) = 0.73$, $p = .53$
Emergency department	74 (68) [3%]	28 (27) [5%]	9 (11) [3%]	54 (26) [7%]	41 (20) [5%]	$F(3, 346) = 0.52$, $p = .67$
Day treatment - mental health care	827 (264) [25%]	642 (244) [24%]	1102 (338) [33%]	1542 (547) [28%]	1029 (186) [27%]	$F(3, 339) = 1.14$, $p = .33$
Hospitalization - mental health care	2557 (1105) [14%]	4385 (1758) [19%]	4417 (1908) [20%]	9158 (2524) [30%]	5150 (967) [21%]	$F(3, 338) = 2.24$, $p = .08$
Hospitalization – hospital	528 (282) [15%]	1537 (589) [19%]	1923 (900) [19%]	2836 (952) [27%]	1711 (370) [20%]	$F(3, 338) = 1.71$, $p = .16$
Total health care costs	**7722 (1535)**	**10215 (2281)**	**12705 (2808)**	**17820 (3349)**	**12135 (1316)**	$F(3, 340) = 2.81$, $p = .04$
Non-health care costs						
Absenteeism	1456 (388) [38%]	2602 (690) [49%]	1925 (514) [39%]	2117 (518) [43%]	2029 (271) [42%]	$F(3, 328) = 0.78$, $p = .51$
Presenteeism	3142 (630) [59%]	5122 (937) [77%]	1968 (437) [62%]	2588 (567) [68%]	3216 (342) [66%]	$F(3, 333) = 4.16$, $p = .01$
Substitution of unpaid work	4421 (870) [79%]	6042 (1183) [82%]	7022 (1489) [88%]	5954 (1139) [79%]	5858 (586) [82%]	$F(3, 337) = 0.80$, $p = .49$
Total societal costs	**16741 (2023)**	**23980 (3277)**	**23620 (3365)**	**28479 (3736)**	**23238 (1612)**	$F(3, 340) = 2.34$, $p = .07$

SE=Standard error.
[a] Psychologist, psychotherapist or psychiatrist.

Cost-effectiveness

Cost-utility acceptability curves are presented in Figure 1. For values of the WTP for one additional QALY between €0 and €100,000, offering the Featback only condition had the highest probability of being efficient for the four alternatives (66%-86%). In other words, between the four conditions, Featback only had the highest probability of having the largest INB across the 1000 bootstrap samples, regardless of the WTP. At very high WTP values, the probability of the combination of Featback with expert-patient support to be efficient compared to the alternatives increased, but still did not exceed the Featback only condition. The care-as-usual condition had a probability of up to 13% of being optimal compared to the alternatives for all WTP values. This probability was around 1% across all WTP values for the expert-patient support only condition.

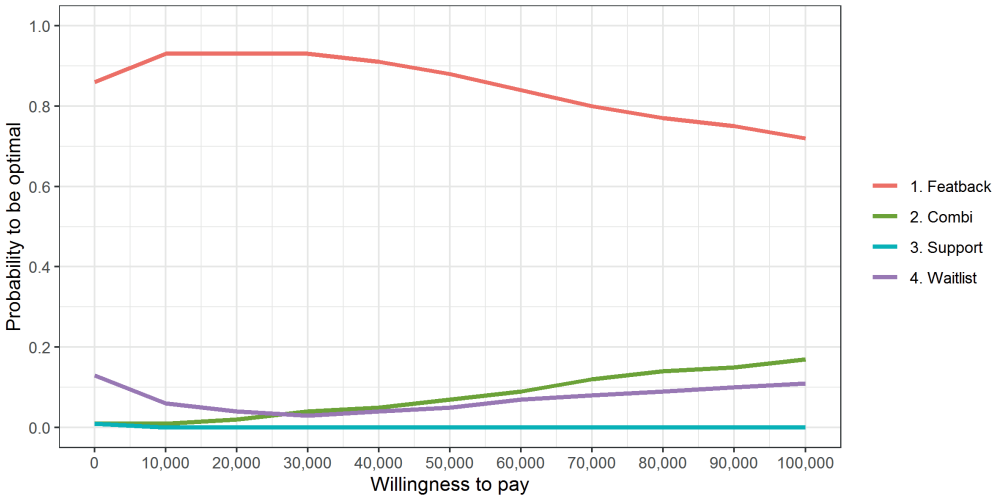
Figure 1. Cost-utility acceptability curves with EQ-5D QALYs for the four study conditions derived from 1000 bootstrap samples



Sensitivity analyses

As can be deduced from Figure 2, results were highly similar for cost-capability acceptability curves, where gains for a particular condition were measured as capability values as assessed with the ICECAP-A. Specifically, the Featback only condition had the highest probability of being efficient compared to the other three conditions across all WTP values (72%-86%). A second sensitivity analysis using the visual analogue scale of the EQ-5D-5L to assess QALYs also showed results comparable to the main analysis. Third, cost-utility acceptability curves of the EQ-5D-5L with direct health care costs only showed the Featback only condition to have the highest probability of being efficient (51%-78%) compared to the three alternatives for WTP values of €60,000 or less. For WTP values between €70,000 and €100,000, Featback

Figure 2. Cost-capability acceptability curves with ICECAP-A capability values for the four study conditions derived from 1000 bootstrap samples



with expert-patient support had the highest probability of being efficient among the four conditions (52%-57%). Fourth, when baseline values of the EQ-5D-5L were estimated using equipercentile mapping with the PHQ-4, an almost identical pattern to the main analysis emerged with the Featback only condition having the highest probability to be efficient compared to the three other conditions across all WTP values (67%-95%). Cost-effectiveness acceptability curves of the last three sensitivity analyses can be found in Appendix F.4.

Discussion

In the current study, an economic evaluation based on a randomized controlled trial covering a period of 14 months was conducted comparing (1) a fully automated internet-based intervention 'Featback', (2) online support by expert patients via mail and chat and (3) the combination of these to (4) care as usual for people with eating disorders. Primarily, from a societal perspective, the Featback intervention had the highest probability of being efficient for a wide range of WTP values compared to the three other conditions. Secondly, expert-patient support alone and care as usual had very low probabilities of being efficient compared to the alternatives over the whole range of explored WTP values. Lastly, the combination of Featback and expert-patient support was more efficient than a care-as-usual condition for WTP values over €20,000, but less than Featback alone. The results suggest that, between the four investigated conditions, Featback is the intervention of choice from a (societal) economic perspective. Despite severe and long-lasting symptoms, 43% of participants in the sample never received treatment for their eating problems, demonstrating the potential of internet-based interventions to reach an underserved population. Notably, as Featback and

expert-patient support are brief interventions, it might be that their impact on quality of life or health care and societal costs is more distinct for people with less severe symptoms or at the beginning stages of their eating disorder, but it proved difficult to reach this group. Furthermore, while around 97% of the Dutch population older than 12 years has internet access, some individuals with eating disorder symptoms cannot be reached through internet or find it challenging to work with and require a different approach. However, implementing the unguided Featback intervention could help to reach individuals with an eating disorder who currently do not receive appropriate care and is likely to lead to similar or slightly better quality of life while reducing costs for society, compared to not implementing it.

The findings are in line with the two studies mentioned in the introduction, cautiously indicating online interventions for eating disorders to be cost-effective compared to care as usual (Akers et al., 2021; Kass et al., 2017). A previous trial concerning Featback indicated that Featback with or without guidance from a psychologist had higher probabilities of being efficient compared to a care-as-usual condition (Aardoom, Dingemans, Spinhoven, et al., 2016). In the current trial, the automed feedback messages were improved and more personalized towards users and support from a psychologist was replaced with expert-patient support, in an attempt to increase the effectiveness of and satisfaction with the interventions. Unexpectedly, conditions with expert-patient guidance were less efficicent than Featback alone for all acceptable WTP values. The contrast with the previous trial, where Featback was equally efficient with and without therapist support, might partly be explained by the improvements to the automated monitoring system, possibly increasing its effectiveness. However, this is unlikely given that the combination condition also included the improved monitoring system, but was outperformed by Featback alone. Another explanation of the favorable probabilities of the Featback only condition in the cost-utility acceptability curves were the (non-significantly) lower costs for several categories, resulting in non-significantly lower total costs. Mainly, a low percentage of people in the Featback condition was hospitalized compared to the other conditions. The finding suggests that brief weekly monitoring and feedback messages can prevent hospitalization. Self-monitoring has been found to be important in preventing psychiatric hospitalization (Ådnanes et al., 2020). However, if the Featback monitoring system would have this effect, lower hospitalization rates in the combination condition would also be expected. Therefore, other explanations why the Featback condition had favorable results, such as chance, cannot be ruled out. Looking across mental disorders, adding guidance appears to increase the probability of being cost-effective compared to unguided alternatives (Donker et al., 2015). However, given the scarcity of evidence, more research directly comparing guided and unguided internet-based interventions is required for decisive conclusions.

More generally, an increasing number of studies is substantiating the claim that internet interventions are likely to be cost-effective compared to care as usual for a number of mental disorders(Ahern et al., 2018; Donker et al., 2015; Hedman et al., 2012; Paganini et al., 2018), but evidence is mixed (Kolovos et al., 2018). The systematic reviews and meta-analysis highlight the need to continue economic evaluation research, since heterogeneity in intervention content and application of guidance make it difficult to reach definitive conclusions. For eating disorder treatment, current evidence indicates that online self-monitoring is likely an efficient alternative to usual care, while more information is needed on whether and how to add guidance. Concordantly, internet-based interventions such as Featback have the

potential to help individuals currently not reached by traditional treatment options, while being worth the investment.

Strengths and limitations

The current study had several strengths, including a 12-month follow-up period with a societal perspective, a large sample size, several sensitivity analyses and multiple imputation procedures to handle missing data. Some limitations can also be noted. First, baseline values of quality of life and wellbeing questionnaires were unavailable. Although a sensitivity analysis with estimated baseline scores produced similar results, the missing values may have led to a slight underestimation of the QALYs and capability values in the active interventions. Relatedly, it could be worthwhile to study which generic preference-based measures are sensitive in eating disorder populations, as both the EQ-5D and ICECAP might be limited in this regard. A second limitation pertains to missing data due to dropout of the current sample. While missing data were handled adequately, imputing cost data was challenging since for some categories only a small percentage of people incurred relatively high costs while others incurred no costs. Thirdly, although many costs were accounted for, assessed direct and non-health care costs were not exhaustive. For example, medication costs and costs attributable to alleviating symptoms were not captured, which may have influenced results slightly. Lastly, cost data were based on self-reported health care visits and work productivity over a period of 4 weeks. This may have introduced recall bias.

Conclusion

A brief fully automated internet-based self-help program (Featback) for eating disorders was found to be efficient compared to care as usual. Results suggest that such interventions may be especially valuable as a first step in a stepped-care model of eating disorder treatment, as it is preferable over a do-nothing approach. Implementing highly scalable and low-threshold interventions such as Featback would not only benefit individuals suffering from an eating disorder, but society as a whole. Mental health professionals and researchers would profit from further investigating how to widely disseminate such interventions to optimize the potential benefits, both in terms of effects and costs. Furthermore, subsidy providers, policy makers and health insurancies should consider wider funding to make installment of evidence-based internet interventions for eating disorders possible, as they appear to be worth the investment.

Chapter 8

Discussion

Discussion

The previous chapters can be read separately but are part of an overarching theme: bridging the large treatment gap that currently exists for individuals with eating disorders with eHealth technology. The research regarding effectiveness and cost-effectiveness in the current dissertation aims to inform on whether, why and how to implement internet-based interventions for eating disorders. In order to reach this goal, research regarding the effectiveness of such interventions is important to determine whether individuals benefit from them and experience improvement in eating disorders symptoms or other outcomes. Additionally, cost-effectiveness research builds towards the main aim by looking at the benefits of internet-based interventions on a societal level: do such interventions reduce costs compared to alternative courses of action and are they good value for money? The current chapter summarizes the main findings of the separate chapters, after which implications, limitations and future directions are discussed.

Summary of main findings

In **chapter 2**, the design of a randomized controlled trial investigating Featback and online expert-patient support for eating disorders was presented. The research aimed at replicating and extending previous findings on Featback (Aardoom, Dingemans, Spinhoven, et al., 2016). Featback is a fully automated internet-based self-help program. After registering, for which only a working email address is necessary, users receive a weekly email over a period of 8 weeks. The email includes a link to a questionnaire with four 4-point questions on eating disorder related symptoms (i.e., weight and shape concern, restrictive food intake, binge eating and compensatory behaviors). After completing the monitoring questionnaire, users receive a supportive feedback message (on average 384 words) that matches the answers, which an algorithm picks from a database with over 1250 handwritten messages. The algorithm also takes the answers of the previous monitoring questionnaire into account, so the feedback messages can contain information on improvement or deterioration. Feedback messages include a summary of self-reported eating problems and changes compared to the previous week, psychoeducation, and guidance on how to counter eating disorder related symptoms. Apart from Featback, expert-patient support was incorporated in the randomized controlled trial design. Expert patients are recovered individuals that have had a lived experience of an eating disorder and may be more approachable by individuals currently suffering from an eating disorder, because of a shared background (Rohrbach et al., 2019). This might make them especially suitable in low-threshold contexts, such as anonymous internet-based interventions. Expert-patient support was offered online in the form of weekly email or chat sessions. Expert patients received training on providing online support and how to use their experiences to help others. The two internet-based interventions, Featback and expert-patient support, and their combination, were hypothesized to have the potential to reduce eating disorder symptoms, but also to reach individuals with eating disorder related problems who currently do not receive appropriate care. To investigate the effectiveness and cost-effectiveness of Featback and expert-patient support, participants aged 16 years or older with at least mild eating disorder symptoms were included. Mild eating disorder symptoms

were defined as a score of 52 or higher on the Weight Concerns Scale (Killen et al., 1993) or, as reported on the Short Evaluation of Eating Disorders (Bauer et al., 2005), a body mass index of 18.5 or lower or at least weekly binge eating episodes or compensatory behaviors in the past four weeks. Participants were randomly assigned to four conditions, being (1) Featback, (2) the combination of Featback and chat or email support from an expert patient, (3) chat or email support from an expert patient, and (4) a waiting list control condition. Participants in all conditions were allowed to make use of other interventions and medication, so the waiting list control condition can be regarded as care as usual for eating disorders. The intervention period was 8 weeks, after which participants were followed for a period of one year. Primarily, the three active interventions were compared to a waiting list in the extent to which they could reduce eating disorder symptoms, as measured by the Eating Disorder Examination Questionnaire (Fairburn & Beglin, 2008). It was also investigated whether the combination condition was superior to Featback and expert-patient support separately, and how Featback and expert-patient support compared to each other. The cost-effectiveness of the three active interventions against the waiting list control condition was also investigated.

The results on the effectiveness of the investigated internet-based interventions in this randomized controlled trial were presented in **chapter 3**. In total, 355 participants enrolled in the study, 43% of which indicated never to have received treatment for their eating disorder, while the average duration of the eating disorder across the sample was more than 10 years. Study dropout rates were acceptable, with the respective number of participants completing assessments at baseline, post intervention, and 3, 6, 9 and 12 month follow up being 355 (100%; T0), 280 (79%; T1), 252 (71%; T2), 244 (69%; T3), 233 (66%; T4) en 242 (68%; T5). Featback only, Featback plus online expert-patient support and online expert-patient support only were effective in reducing eating disorder symptoms after eight weeks compared to a waiting list control condition. On the long term, the difference between the participants allocated to the waiting list control condition and the other conditions disappeared, indicating no long-term differences between the conditions regarding eating disorder symptoms. Surprisingly, the three active interventions were all found to be equally effective. Email and chat sessions of expert patients were compared with psychologist sessions from the previous trial (Aardoom, Dingemans, Spinhoven, et al., 2016). The sessions could reliably be distinguished by blinded master level psychology students, indicating that expert-patient and psychologist support are distinct interventions. Differences concerned expert patients sharing personal experiences, which psychologists never did, and psychologists using a broader pallet of interventions (e.g., stimulating reflection, confrontation, challenging cognitions, etc.) compared to expert patients. In the current study, participants who had the option to receive weekly expert-patient support were more satisfied with the intervention, compared to participants who only had access to the automated Featback intervention. Specifically, participants graded (on a scale from 1 'not at all satisfied' to 10 'extremely satisfied') Featback, expert-patient support and their combination with a 5.8, 7.4 and 7.1 respectively. This indicates that receiving guidance is rated as more satisfying by users, but this does not necessarily lead to increases in effectiveness. Interestingly, despite lower satisfaction ratings, participants tended to complete more Featback monitoring sessions compared to expert-patient support sessions. Participants used an average of 6.5 ($SD = 2.1$) monitoring sessions in the Featback only condition and 5.6 ($SD = 2.7$) monitoring sessions in the com-

bination condition, whereas on average 4.4 ($SD = 3.1$) and 3.6 ($SD = 2.9$) support sessions were used in the expert-patient support only and combination condition. The amount of completed sessions was not associated with the effectiveness of the intervention. Secondary outcome measures that were investigated, including symptoms of anxiety and depression, self-efficacy and experienced social support, yielded no differences between the four conditions. Regarding help-seeking behaviors, 33 participants indicated that the intervention they received stimulated them to get professional help. However, no statistically significant differences between conditions were found in the number of participants that initiated professional help (i.e., had never received eating disorder treatment before the study and did receive such treatment at some point during the study). Taking all the results together, it seems that a low-threshold internet-based intervention like Featback and expert-patient support can help to reduce eating disorder symptoms, at least on the short term. Since 43% of the sample had never received eating disorder related treatment, it appears that this intervention can reach individuals who are not reached by other forms of treatment and foresees in a currently unmet need.

Chapter 4 described a systematic review and meta-analysis of studies investigating the cost-effectiveness of internet-based interventions for mental disorders compared to usual care in society. It provided an overview on cost-effectiveness research in the area of internet interventions for mental disorders and informed on whether such innovative interventions might help to treat individuals with a mental disorder in an efficient way. Specifically, scientific literature databases were searched to find studies describing (1) randomized controlled trials that (2) included participants with symptoms of mental disorders, (3) investigated a telephone or internet-based intervention, (4) included a control condition in the form of treatment as usual, psychological placebo, waiting list control or bibliotherapy, (5) reported outcomes on both quality of life and costs and (6) were published in English. The search also covered unpublished data by including trial registries and contacting authors of published study protocols. Data on risk of bias, quality of the economic evaluation, quality-of-life adjusted life years (QALYs) and costs were extracted from included studies and the incremental net benefit (INB) was calculated and pooled. The INB combines the differences between an internet intervention and a control condition in terms of both the effectiveness (i.e., improvements in quality of life) and their costs into one single monetary value. A positive INB, then, indicates a balance between effects and costs in favor of the internet intervention rather than the control condition and vice versa. The search led to the inclusion of 37 studies with a total of 14,946 participants. The studies investigated depression ($N = 16$), anxiety ($N = 7$), alcohol or substance abuse ($N = 5$), depression and anxiety simultaneously ($N = 5$), obsessive compulsive disorder ($N = 2$), post-traumatic stress disorder ($N = 1$), and eating disorders ($N = 1$). The evaluated internet interventions mostly consisted of online modules based on principles of cognitive-behavioral therapy, but could also be text messaging, web-based games or telephone support. Some form of human support, such as written feedback on exercises or telephone calls, was available for most interventions ($N = 27$), but a minority was fully automated ($N = 10$). A very small, but statistically significant difference, was found between internet-based interventions and control conditions regarding the extent to which they improved quality of life, in favor of the internet-based interventions. Internet interventions and control conditions were found to be equally costly.

This was true for studies that included only costs directly related to the intervention plus costs related to health care service use (i.e., health care perspective), as well as for studies that included these costs plus non-health care costs, such as absence from work and reduced productivity at work because of complaints related to the mental disorder being studied (i.e., societal perspective). Nevertheless, the pooled INB indicating the cost-effectiveness of internet-based interventions compared to control conditions was positive ($255, 95% CI $91; $419). This suggests that internet interventions for mental disorders are likely to be cost-effective compared to a waiting-list or care-as-usual approach. Several explorative moderator analyses were conducted to better understand the results. They indicated that interventions aimed at depression and anxiety, and interventions that incorporate some form of guidance, might have higher probabilities of being efficient compared to ones aimed at other mental disorders or unguided ones. Additionally, the perspective of the economic evaluation (health care or societal) influenced whether internet interventions for mental disorders were found to be cost-effective compared to control conditions. Unlike studies with a health care perspective, pooling studies with a societal perspective resulted in a non-significant difference between the comparators. When interpreting the results, it is important to note that the large heterogeneity between studies regarding type of internet intervention, included costs and targeted mental disorder impede definitive conclusions. Nevertheless, the findings highlight the potential economic benefits of internet-based interventions for mental disorders. The used method of pooling cost-effectiveness data in an aggregate-data meta-analysis is new in the area of psychiatry and might help other researchers in pursuing related research questions.

Surprisingly, in the meta-analysis, internet interventions were found to be significantly more effective than control conditions in terms of quality of life improvements, but the difference was very small. It may be that quality-of-life questionnaires that are often used in economic evaluations (e.g., EQ-5D) are not suitable in contexts outside of (somatic) health, such as psychiatry (Mitchell et al., 2017). The ICECAP-A instrument is increasingly proposed as an alternative, because it might be more sensitive to changes in well-being in such contexts (Keeley et al., 2016; National Institute for Health and Care Excellence, 2016). The ICECAP-A measures five capabilities important to one's quality of life: (1) stability - the extent to which someone can feel settled and secure; (2) attachment - the extent to which someone can feel love, friendship, and support; (3) autonomy - the extent to which someone can feel independent; (4) achievement - the extent to which someone can experience achievement and success; (5) enjoyment - the extent to which someone can experience enjoyment and pleasure. Four levels are available for each of the five capabilities, ranging from 1 (not being able to experience a capability at all) to 4 (fully being able to experience a capability). The ICECAP-A attempts to capture the extent to which someone experiences the ability to do what one wishes. In **chapter 5**, psychometric properties of the Dutch version of the ICECAP-A were investigated to clarify its added value in economic evaluations. Specifically, the test-retest reliability of the ICECAP-A was examined to evaluate the extent to which the instrument is prone to measurement errors. Furthermore, the instrument was related to the EQ-5D-5L and a measure of self-efficacy to better understand the concept the ICECAP-A measures. Lastly, the ability of the ICECAP-A to distinguish groups that were hypothesized to differ in terms of quality of life was studied. Results on test-retest reliability were based

on 252 participants and indicated acceptable measurement error. Other results were based on the answers of 1002 participants. The ICECAP-A correlated moderately to strongly with the EQ-5D-5L, indicating that there is overlap between the two instruments, but they do not seem to measure the same concept. The concept captured by the ICECAP-A correlated strongly with self-efficacy, which encompasses the extent to which individuals believe in their own capability and mastery over life. Additionally, the ICECAP-A was able to differentiate groups based on, among others, happiness ratings, the presence of a chronic illness, self-efficacy ratings and employment status. Being able to distinguish groups with varying levels of quality-of-life status is, arguably, an important feature when measuring changes in people over time or after having received a certain treatment. In summary, the ICECAP-A appears to capture a concept that correlates with health status (as measured by the EQ-5D-5L), but more closely relates to self-efficacy and well-being. Therefore, it can complement other generic health questionnaires in economic evaluations, especially when attempting to capture the benefits of interventions outside hospital contexts, such as elderly care, social care, psychiatry, chronic illness and general living conditions.

In **chapter 6**, a tariff was developed to be able to reliably use the ICECAP-A in economic evaluations in the Netherlands. Specifically, a best-worst scaling task was constructed to be able to compare individual ICECAP-A domains (i.e., stability, attachment, autonomy, achievement and enjoyment) and levels (ranging from 1 to 4) with each other. A sample of 1002 participants representative of the Dutch general population was recruited by a market research agency. Based on the participants' answers on the best-worst scaling task, the relative importance of each of the ICECAP-A items to one's quality of life could be distilled. Results indicated that all five domains are valuable contributors to one's quality of life for the general Dutch population, but that stability, attachment and enjoyment are especially important. Moreover, improving low levels of a domain (e.g., going from level 1 to level 2) is considered more valuable than improving already moderate levels of a domain (e.g., going from level 3 to level 4). This suggests that helping people with low capabilities, rather than those with high capabilities, would result in larger gains in well-being for society. Interestingly, when comparing the ICECAP-A tariff for the Dutch and UK general population, it seems the Dutch population values enjoyment somewhat more and achievement less than their UK counterparts. Consequently, interventions that increase the ability to enjoy life may have a slightly greater impact on well-being in the Netherlands than in the United Kingdom. Now that a tariff for the ICECAP-A for the Dutch general population has been developed, the instrument can be used in economic evaluations in the Netherlands and might complement other instruments such as the EQ-5D.

After looking more broadly at cost-effectiveness research on e-mental health interventions and possible improvements in measuring benefits of interventions by incorporating the ICECAP-A, **chapter 7** returns to internet-based interventions for eating disorders. This chapter extends the findings on Featback, expert-patient support and their combination for eating disorders by looking at their cost-effectiveness compared to usual care (waitlist). Participants in the four conditions were inquired after their quality of life, measured both by the EQ-5D-5L and ICECAP-A. Dutch tariffs for the two instruments were used to calculate quality-of-life adjusted life years (QALYs) and capability values over the 14-month study period. Fur-

thermore, costs related to the intervention (e.g., personnel), health care service use related costs and non-health care costs (e.g., absence from work or reduced efficiency at work) were estimated over the study duration. Finally, to compare the cost-effectiveness of the different conditions, cost-utility acceptability curves were constructed. These curves show the probability of one condition to be cost-effective (i.e., efficient) compared to another, across a range of willingness to pay (WTP) values. WTP values indicate society's willingness to pay for one extra year lived in perfect health (i.e., 1 QALY). The higher the WTP value the more important differences in QALYs between the four conditions become when evaluating cost-effectiveness. To accommodate all relevant WTP values, the study explored values ranging from €0 to €100,000 per QALY. Results on QALYs and capability values showed improvements over time, with only minor and non-significant differences between the four conditions (i.e., Featback, combination of Featback and expert-patient support, expert-patient support and waiting list control). Similarly, no significant, but still noteworthy, differences between the four investigated groups regarding health care expenses and societal costs over a period of 14 months were found. When looking at used health care services during the study duration, almost all participants saw their general practitioner (92%), a majority visited a psychologist, psychotherapist or psychiatrist (85%) and hospitalization in the hospital or a mental health institution was common (31%). Costs related to health care service use were lowest in the Featback only condition, mainly because hospitalization rates tended to be lower in this condition, but differences were not statistically significant. Societal costs, including intervention costs, health care costs and non-health care costs, in the Featback, combination, support and waiting list condition were €16,741, €23,980, €23,620 and €28,479 respectively. Looking at the cost-utility acceptability curves, providing Featback without expert-patient support had the highest chance to be efficient among the alternatives, regardless of how much society was willing to pay per QALY. Comparable results were found when ICECAP-A deduced capability scores were used, suggesting that in the current sample there was no clear preference for using the EQ-5D-5L or the ICECAP-A. Overall, the findings propose that, between the four investigated conditions, Featback is the intervention of choice from a (societal) economic perspective.

Implications

Bridging the treatment gap

The findings from this dissertation provide important implications and recommendations for policy, service delivery, and clinical practice. Most importantly, results from the conducted studies confirm the potential of internet-based interventions for eating disorders to reach the large of group individuals in society who suffer from eating disorder related problems but who do not get appropriate care. Even people with long standing and severe eating disorder symptoms can benefit from easy-access brief online interventions and experience a degree of symptom reduction. Reaching individuals who have never received (appropriate) care and guiding them to treatment for their eating disorder will go a long way in improving effectiveness of eating disorder treatment overall (Moessner & Bauer, 2017). Low-threshold and highly scalable internet-based interventions such as Featback and expert-patient support

can be an important tool to bridge the large treatment gap that exists for eating disorders.

From a (societal) economic perspective it would also make sense to implement internet-based interventions. Indeed, both the economic analysis of Featback and expert-patient support and the conducted meta-analysis suggest such interventions to be an efficient use of resources when compared to care as usual. In other words, they provide quality-of-life improvements at a relatively cheap rate. Consequently, widely disseminating internet-interventions for eating disorders will likely be beneficial for individuals with eating disorder symptoms and for society. Policy makers, clinicians and researchers should consider strategies for achieving a more widespread adoption of interventions such as Featback. Furthermore, subsidy providers, policy makers, municipalities, and health insurance companies are encouraged to provide an adequate financial framework, to stimulate implementation of evidence-based internet interventions for eating disorders, as they appear to be worth the investment or even save costs.

Guidance

There does not appear to be a clear answer on whether and how to include guidance to internet-based interventions. Across mental disorders, adding guidance to eHealth interventions seems to yield an increase in effectiveness (Baumeister et al., 2014). In the meta-analysis by Baumeister et al. (2014), guidance was applied differently in various studies and was mostly provided by psychologists. The evidence for the effectiveness of guidance in internet-based interventions for eating disorders is mixed (Barakat et al., 2019; Yim & Schmidt, 2019b). Accordingly, the trial described in this dissertation (chapter 3) did not find an added effect of guidance to Featback when provided by expert patients. The finding that Featback only, expert-patient support only and their combination were equally effective in reducing eating disorder symptoms might be because they offer a similar intervention. That is, all three interventions help to make users aware of and engage their eating disorder, while the way in which such an intervention is conveyed appears to be of lesser importance. Nevertheless, it is surprising that expert-patient support did not add to the effectiveness of Featback alone. An explanation might be that expert patients are less effective than other health professionals in providing online support. However, this is unlikely given that a previous large trial on Featback also found no added effect of guidance when provided by a psychologist (Aardoom, Dingemans, Spinhoven, et al., 2016). Still, email and chat sessions of expert patients from the current trial (chapter 3) and psychologist support from the previous trial could reliably be distinguished, suggesting that they are distinctly different interventions. The most apparent differences were that expert patients shared their own experiences during the sessions and psychologists used a broader pallet of interventions. It was thought that the shared background of participants and expert patients would result in a greater perceived similarity (i.e., the extent to which participants felt similar to their supporter), stimulating bonding and improving self-efficacy. No proof for this was found. In fact, perceived similarity ratings were surprisingly low. The questionnaire to assess perceived similarity was sent out in week three of the intervention. Perhaps three weeks is too short to establish rapport. Another explanation is that participants might focus on differences between themselves and their assigned expert patient to reduce feelings of failure and a negative self-image. For example, acknowledging similarities to recovered individuals while not

being recovered oneself might lead to cognitive dissonance. Such upward comparison has been found to decrease self-evaluation (Collins, 1996; Wayment et al., 2020). Perceiving to be different, then, might make it easier to accept not being recovered. Furthermore, findings indicated lower intervention completion rates for expert-patient support compared to the automated monitoring intervention, suggesting that participants experienced barriers to make appointments with their expert patient. Contrary to the findings in our randomized controlled trial (chapter 3), across mental disorders, intervention adherence seems to be higher when guidance is added (Musiat et al., 2022). The discrepancy may be explained by the operationalization of guidance. Indeed, in many trials guidance comprises providing some form of feedback on an otherwise automated intervention, like completing online modules. In the trial described in this dissertation, expert-patient support involved email or chat sessions as separate intervention. These two conceptualizations of guidance, as part of an automated intervention or as a separate intervention, make it difficult to compare the existing literature and the conducted trial regarding the influence of guidance on adherence. Regardless, the relatively low number of completed expert-patient support sessions is surprising given that participants were more satisfied with the intervention when they had the option to receive expert-patient support compared to when they could only access the automated monitoring system. Previous research has also indicated that guidance accompanying eHealth interventions is highly valued (Linardon et al., 2021; Yim & Schmidt, 2019a), though it does not necessarily result in other benefits (Berger et al., 2011; Ciuca et al., 2018).

Taken together, it would make sense to offer Featback without any guidance, as it is highly scalable, requires low maintenance and was found to have the highest probabilities to be cost-effective while being equally effective compared to Featback with expert-patient support and expert-patient support alone for eating disorders. However, determining user preferences and users' ideas about intervention fidelity even before commencement of the interventions and examining how they influence uptake is needed to further substantiate implementation decisions.

Dissemination strategies

Apart from deliberating whether adding guidance helps with the dissemination of internet-based interventions such as Featback, other dissemination strategies should be considered. Indeed, the sample in our randomized controlled trial had a long average duration of illness and high levels of eating disorder symptomatology. Attempts were made to recruit a younger sample with beginning eating disorder problems, for instance by advertising at a health and lifestyle website mostly targeting adolescent girls. Reaching individuals with an eating disorder at an early stage is important, as the chance of recovery appears to be higher for those with a shorter illness duration (Vall & Wade, 2015). However, those expressing interest in participating in the study tended to find the study via the Proud2Bme community and had been struggling with a serious eating disorder for quite some time. Varying dissemination strategies (e.g., via high schools, websites, social media and flyers) might help to reach a large and diverse population (Bauer et al., 2019), increasing the overall impact.

Cost-effectiveness

Finally, researchers and clinicians working in the field of internet-based interventions for mental disorders are encouraged not to shy away from cost-effectiveness research. Economic evaluation is a useful tool to complement effectiveness research, as it adds to the understanding of potential benefits of an intervention compared to other relevant courses of action. Consequently, more information is available to choose between different treatment strategies, which will ultimately result in effective care at a fair price. Cost-effectiveness research might be especially important in the area of eHealth, because automating certain aspects of an intervention could lead to high scalability and low intervention costs, making it an efficient alternative compared to other courses of action. However, a good understanding of what cost-effectiveness research entails is necessary to critically evaluate its added value and limitations. For example, the often used QALYs might be limited in their extent to capture benefits of an intervention, especially when evaluating an intervention aimed at improving quality of life beyond (physical) health (Pietersma et al., 2013). In such cases, it might be beneficial to complement established health questionnaires that capture QALYs with other instruments, such as the ICECAP. The ICECAP seems to capture a concept that is related to health as captured by the EQ-5D-5L, but approaches well-being more broadly (Afentou & Kinghorn, 2020; Al-Janabi et al., 2013; Rohrbach, Dingemans, Essers, et al., 2022). Interestingly, no noteworthy differences emerged when using the EQ-5D-5L or the ICECAP-A in the economic evaluation concerning Featback and expert-patient support, suggesting the ICECAP-A was not as sensitive as expected in the recruited sample. This is not in line with previous research suggesting that the ICECAP-A might be more sensitive in capturing quality-of-life changes of individuals with a psychiatric disorder like depression (Mitchell et al., 2017). Results indicated that individuals in the conducted trial described in this dissertation (chapter 7), on average, experienced a reduction in eating disorder symptoms, but no major changes in quality of life as assessed by the EQ-5D-5L, its visual analogue scale and the ICECAP-A. Since three separate measures indicated only minor group differences, perhaps the findings are not attributable to a lack in instrument sensitivity, but rather to the interventions having no substantial effect on quality of life. A second explanation of why the ICECAP-A and the EQ-5D-5L showed similar results might be that eating disorders, in comparison to depression for example, involve a considerable somatic component. For eating disorders specifically, then, the EQ-5D-5L might be a sensitive tool for measuring quality of life, even if this is not the case for other mental disorders. All in all, more investigation is needed to determine which generic preference-based measures are sensitive in eating disorder populations. Nevertheless, using the ICECAP instead of other quality-of-life measures has in some cases been found to lead to different recommendations concerning resource allocation (Kiadaliri et al., 2015). Therefore, it is recommended in the Netherlands to add the instrument when conducting cost-effectiveness research of interventions in the area outside (somatic) health, such as elderly care, chronic illness, general livability and psychiatry (Zorginstituut Nederland, 2015). A Dutch general population tariff is now available, so that the ICECAP-A can be used reliably in Dutch samples. Apart from the ICECAP-A and quality-of-life measures like the EQ-5D, other instruments might be appropriate to use in economic evaluations, depending on the context. For example, Van Krugten et al. (2022) have recently developed the Mental Health Quality of Life questionnaire (MHQoL), a

quality-of-life questionnaire for economic evaluations in the area of mental health specifically. The MHQoL contains seven questions on self-image, independence, mood, relations, daily activities, physical health and optimism about the future. A first study indicated promising psychometric properties (Van Krugten et al., 2022). While additional corroboration is needed to determine its added value, such a measure might further enhance the precision of evaluating intervention benefits in the area of mental health, including internet-based interventions. In summary, the field of cost-effectiveness research for eHealth interventions for mental health is progressing and helps to inform clinicians, subsidy providers and policy makers on the value of such interventions.

Strengths and limitations

The conclusions from the randomized controlled trial on Featback and expert-patient support were based on a large sample size of 355 participants. The participants were recruited via online platforms in such a way that the final sample was representative of the intended users of the online interventions. Therefore, the results can reliably be generalized to real-world settings. It must be noted, however, that generalization to younger individuals with a shorter illness duration is limited (addressed later in this section). Missing data were multiply imputed, which is preferred over other methods of dealing with missing data (Van Ginkel et al., 2020). This reduced the influence of missing data on the final results, and can be considered a strength. Another major strength of the current dissertation was that results were not limited to effectiveness research, but also included a thorough investigation of the cost-effectiveness of internet-based interventions. This improved the understanding of the value of such interventions from an economic perspective, which is important when making decisions on financing and implementing them.

Several limitations can also be noted. First, intervention and study dropout occurred in the randomized controlled trial. Intervention completers were defined as participants who used at least five out of eight monitoring sessions, five out of eight expert-patient support sessions or both, depending on whether they were allocated to the Featback, expert-patient support or combination condition. Across the three active conditions there were 156 (59%) completers: 74 (84%) in the Featback, 48 (55%) in the expert-patient support and 34 (38%) in the combination condition. Consequently, participants likely experienced barriers to make full use of the interventions, especially with regard to the expert-patient support. Perhaps pro-actively scheduling appointments with a supporter through an online system was too much work or easily forgotten. In comparison, the automated self-help program required less time and effort, as participants only had to click a link in an email that was received every week. In general, intervention dropout appears to be common in internet-based interventions, ranging between 9% and 47% (Dölemeyer et al., 2013). Apart from intervention dropout, study dropout was also a concern. About one-third of the participants dropped out of the study over the course of 14 months. Such study dropout can affect results. Nevertheless, the dropout was similar to the 21% dropout at post intervention found in a recent meta-analysis (Linardon et al., 2020) and substantially lower than in the previous trial (Aardoom, Dingemans, Spinhoven, et al., 2016). This might partly be ascribed to improvements in the software (resulting in less technological issues with Featback), scheduling sessions

and the online questionnaires. Additionally, reminders and communication were optimized and personalized (e.g., sending an email after completing an assessment including words of appreciation and information on when to expect the next questionnaire), with the goal of retaining participants.

A second limitation pertains to the recruited sample of the randomized controlled trial, which, as mentioned before, mostly included participants with severe symptoms and a long duration of eating-related problems. The current sample is therefore hard to generalize to individuals with less severe eating disorder symptoms, for whom such brief low-threshold interventions may be especially useful.

Thirdly, outcomes in the randomized controlled trial were exclusively determined using self-report measures. This limited the diagnostic precision of outcomes and might have introduced recall bias. Nevertheless, relying solely on self-report measures was necessary to maintain the low-threshold character of the interventions and making the research easily accessible for participants.

A fourth limitation worth mentioning concerns the conducted meta-analysis (chapter 4). Specifically, it is hard to precisely value the outcomes of the meta-analyses pooling 37 studies, because the studies show substantial heterogeneity. Heterogeneity between studies is a common problem in meta-analyses (Cuijpers, 2016), but might be especially relevant in meta-analyses on cost-effectiveness data, as differences do not only relate to the design, content of interventions and effectiveness outcomes, but also to the way economic outcomes are assessed. The concerns were addressed by conducting moderator and sensitivity analyses, which help to better understand the available data, but generalizability to other studies remains limited. Still, meta-analyses of cost-effectiveness studies are arguably important, as singular economic evaluations are often conducted after effectiveness analyses and lack power for cost-effectiveness analyses (Hollingworth et al., 2013). The conducted meta-analysis is an illustration on how pooling cost-effectiveness data can be practically accomplished. This has been achieved before in the area of medicine (Bagepally et al., 2019), but is the first in the area of psychiatry. It paves the way for other researchers in the area of psychiatry to build on and improve the presented method.

Future directions

Low-threshold interventions like Featback appear to be useful in bridging the eating disorder treatment gap, by reaching individuals who indicated never to have had eating disorder treatment before. However, seeing how participants in the conducted randomized controlled trial were already severely ill, the challenge remains to detect individuals at the beginning stages of their eating disorder. Shifting the attention of all those involved in eating disorder treatment to this task will likely be beneficial in several ways (Moessner & Bauer, 2017), including the realization of cost savings. Apart from conventional dissemination techniques that might help to reach youngsters with eating disorder symptoms, such as handing out flyers in high schools (Bauer et al., 2019), other approaches need to be explored. First, it may be valuable to work with experts in communication and sales, as health professionals and researchers are often not well versed in (non-scientific) mass communication. Second, social media such as Instagram and TikTok have become an increasingly large part of the

lives of teenagers and their use is frequently associated with eating disorder symptoms (Padín et al., 2021). Understanding these platforms and using them to reach and inform those with eating-related problems appears to be an essential step for wide dissemination of internet-based self-help programs for eating disorders. For example, creating and maintaining a channel to promote effective interventions or working together with influencers with a large network might help to reach individuals with eating disorder symptoms at an early stage.

Two large trials on Featback, either with guidance from psychologists (Aardoom, Dingemans, Spinhoven, et al., 2016) or expert patients (chapter 3), found no added effect of guidance on eating disorder symptom reductions. Indeed, apart from higher satisfaction, it appears difficult to specify benefits of adding support to unguided internet-based interventions such as Featback. The results advocate a reappraisal of the role of guidance in internet-based interventions for eating disorders. The unguided version of Featback is highly scalable and likely to be more efficient than its guided form. Abandoning guidance in the form of psychologist or expert-patient support altogether and focusing on improving the automated part of the intervention, then, seems sensible. Nevertheless, investigating how adding (expert-patient) support to Featback influences uptake might still be an interesting research avenue. Additionally, future trials on internet-based interventions for eating disorder are recommended to include an unguided version of the intervention of interest to investigate whether human support is necessary for effectiveness.

Another venerable pursuit is the continuation of high-quality cost-effectiveness research. A recommendation for researchers would be to always include an economic evaluation when planning future randomized controlled trials on eHealth interventions and to involve a health economist early in the conceptualization phase. Growing the body of evidence in this field will provide a clearer picture of the economic impact of internet-based interventions for eating disorders. This will inform both clinicians and policy makers on the value of such interventions.

Lastly, future directions may also concern further personalization of internet-based interventions for eating disorders. Technological possibilities become increasingly versatile, making it feasible to have interventions that adapt to the needs of the user. Currently, the majority of interventions are limited in their responsiveness to user input and are often text based (Burger et al., 2020). It seems that the development of an internet-based intervention is still frequently approached as taking an effective (face-to-face) treatment and translating it into an online application. However, the technological innovations at our disposal open up possibilities that might aid in reaching a large audience and developing engaging and effective interventions. When considering to incorporate novel technologies, working together with intended end-users during development might have beneficial effects. Indeed, such a co-creation process engages users from an early stage and identifies actual needs of end-users, arguably leading to effective and persuasive designs that users enjoy engaging with (Alpay et al., 2019). Advancement in areas like, but not limited to, Virtual Reality, machine learning, artificial intelligence and gamification may offer unique opportunities to enhance existing interventions. For example, a chatbot for eating disorder prevention, simulating human conversation, has been developed and tested in a randomized controlled trial and was found to reduce weight and shape concerns and overall eating disorder psychopathology (Fitzsimmons-Craft et al., 2022). Another approach, making use of smartphone technology, is just-in-time (JIT) interventions. In such JIT interventions, data are collected frequently,

for instance using short questionnaires on mood and well-being, or even in real-time (e.g., physical activity and heartrate). These data are then used in (machine-learned) algorithms to predict the optimal timing and content of interventions, messages, or exercises for a specific individual. This approach is uncommon still in eating disorders, but enables highly personalized interventions (Juarascio et al., 2018). The two examples mentioned here serve as illustration of innovative ways to use technology and as inspiration of how existing interventions might be creatively enhanced without requiring new, large research trials that take years to complete. Indeed, there is some evidence to suggest that using gamification, videos or Virtual Agents may be useful in improving mental health and intervention engagement (Abd-Alrazaq et al., 2020; Fleming et al., 2017). Similarly, Barakat et al. (2019) found that using multiple features that address different modalities might have a positive influence on the effectiveness of technology-enhanced eating disorder treatments. The Featback intervention already involves some responsiveness to user input (i.e., personalization), by sending a message to users that is based on the symptoms they report. However, the intervention may become more effective and appealing to use when other features are added, such as a Virtual Agent that sets goals and motivates users to try out the tips in the feedback messages or rewarding intervention usage with badges or unlockable clothing for a virtual character that users see when logging in. Replacing expert-patient support with a chatbot that is available upon demand or sending an appropriate feedback message as a response to passively collected data via the smartphone (i.e., JIT intervention) are other examples. Arguably, such enhancements are most effective when they have a scientific underpinning, are found to be feasible in a co-creation process, and are incorporated in collaboration with technical experts.

Conclusion

Based on the current dissertation and earlier research, it appears low-threshold internet-based interventions for eating disorders, such as Featback, can complement existing treatment options in three ways. First, they have been repeatedly found to be effective in reducing eating disorder symptomatology. Second, such interventions can reach individuals that are currently not reached by other forms of treatment and stimulate them to get professional help. Finally, internet-based interventions are likely to be cost-effective compared to care as usual. Concordantly, implementing highly scalable and easily accessible interventions like Featback likely helps to reduce both the individual and societal burden of eating disorders.

Nederlandse samenvatting

Er is in toenemende mate belangstelling voor het inzetten van internet technologie om de geestelijke gezondheidszorg te verbeteren. Ook voor eetstoornissen, die vaak leiden tot een ernstige verstoring in het functioneren, verminderde kwaliteit van leven of zelfs overlijden, zouden zulke eHealth toepassingen waardevol kunnen zijn. Ondanks de ernst ervaren individuen met een eetstoornis vaak veel schaamte en weerstand tegen het zoeken van passende hulp. In **hoofdstuk 1** van dit proefschrift wordt stil gestaan bij de lange periode tussen het ontstaan van eetstoornis gerelateerde klachten en het zoeken van passende hulp, ook wel de 'behandelkloof' genoemd. Internet interventies kunnen deze behandelkloof voor eetstoornissen helpen verkleinen, door hun laagdrempeligheid, toegankelijkheid en schaalbaarheid. Zulke innovatieve toepassingen beloven niet alleen mensen met eetstoornissymptomen eerder naar passende zorg te leiden, maar kunnen ook helpen bij het verminderen van de klachten. Bovendien hebben ze de potentie tot zorgkostenbesparing, omdat mensen in een eerder stadium van hun eetstoornis worden bereikt, wat een ongunstig beloop met lange zorgtrajecten en meer ziekteverzuim zou kunnen voorkomen. Kortom, bij het onderzoeken van de voordelen van internet interventies voor eetstoornissen is het niet alleen belangrijk om de effectiviteit en het bereik te bestuderen, maar ook de (zorg)kosten ten opzichte van andere behandelopties, zoals face-to-face therapie of niets doen.

Dit proefschrift heeft als doel om meer kennis te vergaren over of en hoe internet interventies een waardevolle toevoeging zijn aan het palet van beschikbare behandelopties voor eetstoornissen. De resultaten zouden kunnen bijdragen aan het bereiken van individuen met een eetprobleem of eetstoornis die momenteel niet bereikt worden. Om dit doel te realiseren lijken laagdrempelige online interventies met passende begeleiding veelbelovend. Specifiek staan in dit proefschrift Featback, een online zelfhulpprogramma voor mensen met eetproblemen of een eetstoornis, en online chat of e-mail ondersteuning door een ervaringsdeskundige centraal. Van deze interventies worden de effectiviteit en kosteneffectiviteit onderzocht. Daarnaast wordt het beschikbare onderzoek naar de kosteneffectiviteit van internet interventies voor mentale stoornissen in het algemeen systematisch samengevat en geanalyseerd. Hieronder volgt een samenvatting per hoofdstuk van dit proefschrift.

In **hoofdstuk 2** werd het design van het gerandomiseerde gecontroleerde onderzoek naar Featback en online ondersteuning door een ervaringsdeskundige gepresenteerd. Featback is een volledig geautomatiseerd internet zelfhulpprogramma. Na registratie, waarvoor alleen een geldig mailadres nodig is, kregen gebruikers een wekelijkse e-mail gedurende een periode

van 8 weken. In de e-mail stond een link naar een vragenlijst met vier vragen over eetstoornis gerelateerde klachten (i.e., zorgen over gewicht en figuur, restrictieve voedsel inname, eetbuien en compensatoir gedrag). Wanneer de monitoringsvragenlijst was ingevuld, kreeg de gebruiker een steunend feedback bericht (gemiddeld 384 woorden) dat paste bij de antwoorden. Het feedback bericht werd door een algoritme gekozen uit een database met meer dan 1250 handgeschreven berichten. Het algoritme onthield ook de antwoorden op de monitoringsvragenlijst van de vorige keer, zodat de berichten informatie konden bevatten over mogelijke verbeteringen of verslechteringen van de eetstoornissymptomen. De inhoud van de berichten besloeg een terugkoppeling van de gerapporteerde klachten, psycho-educatie over de klachten en tips over het omgaan met de genoemde klachten. Naast Featback werd ook chat en e-mail ondersteuning door een ervaringsdeskundige bestudeerd in het onderzoek. Een ervaringsdeskundige had zelf een eetstoornis gehad en was hiervan hersteld. De hypothese was dat ervaringsdeskundigen geschikt zijn voor een laagdrempelige internet interventie, omdat ze makkelijker te benaderen zijn dan andere gezondheidsexperts zoals psychologen en er snel een band met open communicatie kan ontstaan. In het onderzoek vond de e-mail of chat ondersteuning, gegeven door getrainde ervaringsdeskundigen, één keer per week gedurende een periode van 8 weken plaats. Er werd verwacht dat de twee interventies, Featback en ondersteuning door een ervaringsdeskundige, en hun combinatie eetstoornissymptomen konden verminderen, mensen konden bereiken die nog geen passende eetstoornisbehandeling kregen en ook economische voordelen met zich mee kon brengen. Om de effectiviteit en kosteneffectiviteit te onderzoeken werden deelnemers van 16 jaar of ouder met minstens milde eetstoornisklachten willekeurig ingedeeld in vier condities: (1) Featback, (2) de combinatie van Featback en wekelijkse chat of e-mail ondersteuning door een ervaringsdeskundige, (3) chat of e-mail ondersteuning door een ervaringsdeskundige, (4) een wachtlijst controle conditie. Deelnemers in alle condities mochten andere zorg, interventies of medicatie ontvangen, dus de wachtlijst controle conditie kon worden gezien als gebruikelijke zorg voor mensen in de maatschappij met een eetstoornissymptomen. De interventieperiode was 8 weken, waarna deelnemers voor een periode van één jaar werden gevolgd. Er werd gekeken of deelname aan de drie internet interventies leidde tot een sterkere afname in eetstoornissymptomen dan de wachtlijst controle conditie, of de combinatie van Featback en ondersteuning superieur was ten opzichte van de twee individuele interventies afzonderlijk, en hoe Featback en ondersteuning door een ervaringsdeskundige zich ten opzichte van elkaar verhielden. Ook werd de kosteneffectiviteit vergeleken tussen de vier condities.

De resultaten van het gerandomiseerde controleerde onderzoek over de effectiviteit van de onderzochte internet interventies werden in **hoofdstuk 3** gepresenteerd. In totaal deden 355 deelnemers mee aan het onderzoek, waarvan 43% nog nooit eerder een eetstoornisbehandeling had gehad terwijl de gemiddelde duur van eetstoornisklachten meer dan 10 jaar besloeg. Het aantal deelnemers dat bleef meedoen aan de studie was acceptabel met op baseline (T0), post interventie (T1; 8 weken) en 3, 6, 9 en 12 maanden follow-up (T2-T5) respectievelijk 355 (100%), 280 (79%), 252 (71%), 244 (69%), 233 (66%) en 242 (68%) deelnemers. Na 8 weken lieten deelnemers in de drie actieve condities (alleen Featback, Featback plus ondersteuning en alleen ondersteuning) een grotere reductie in eetstoornissymptomen zien dan deelnemers in de controle conditie. Op de lange termijn verdwenen de verschillen tussen de vier condities. Verassend was dat de drie actieve condities even effectief werden bevon-

den. Wel waren deelnemers meer tevreden met de interventies zodra ze de optie hadden om ondersteuning door een ervaringsdeskundige te krijgen. Ondanks de hogere tevredenheidscijfers maakten deelnemers minder gebruik van ondersteuning door een ervaringsdeskundige (gemiddeld 4 sessies gedurende 8 weken) dan van de monitoringsvragenlijsten (gemiddeld 6 Featback berichten gedurende 8 weken). Ook werd er gekeken naar verschillen tussen de vier onderzochte condities op het gebied van symptomen van angst en depressie, zelf-effectiviteit ('self-efficacy'; het gevoel zelf gezette doelen te kunnen bereiken) en sociale steun, maar deze werden niet gevonden. Met betrekking tot hulp zoeken gaven 33 deelnemers aan dat de interventie hen stimuleerde om te starten met professionele hulp. Er werd echter geen statistisch significant verschil gevonden tussen de vier condities in het aantal deelnemers dat professionele hulp initieerde gedurende de onderzoeksperiode. Samengevat lieten de resultaten zien dat laagdrempelige internet interventies zoals Featback en ondersteuning door een ervaringsdeskundige hielpen bij het op korte termijn verminderen van eetstoornissymptomen. Aangezien 43% van de deelnemers nog nooit eerder een eetstoornisbehandeling kreeg, leken deze interventies in een behoefte te voorzien en mensen te kunnen bereiken die niet werden bereikt door andere vormen van behandeling.

In **hoofdstuk 4** werd een systematisch literatuuronderzoek en meta-analyse gepresenteerd over kosteneffectiviteit van internet interventies voor mentale stoornissen vergeleken met gebruikelijke (maatschappelijke) zorg. Er werden 37 studies met in totaal 14,946 deelnemers geïncludeerd. In de studies werden internet interventies onderzocht voor depressie ($n = 16$), angst ($n = 7$), alcohol of drugsmisbruik ($n = 5$), angst en depressie gezamenlijk ($n = 5$), obsessieve-compulsieve stoornis ($n = 2$), posttraumatische stressstoornis ($n = 1$) en eetstoornissen ($n = 1$). De onderzochte internet interventies waren grotendeels online modules gebaseerd op principes van cognitieve gedragstherapie, maar konden ook bestaan uit sms'jes, online educatieve spelletjes of ondersteuning via de telefoon. Bij de meeste interventies was een vorm van menselijke ondersteuning aanwezig zoals geschreven feedback op oefeningen of een telefoongesprek ($n = 27$), maar sommige interventies waren volledig geautomatiseerd ($n = 10$). Internet interventies en controle condities (meestal een wachtlijst of gebruikelijke maatschappelijke zorg) kostten evenveel. Dit ging op voor studies die een gezondheidsperspectief innamen en dus alleen kosten direct gerelateerd aan de interventies en gezondheidszorgkosten meenamen ($n = 15$) alsook voor studies met een maatschappelijk perspectief, die ook kosten gerelateerd aan ziekteverzuim en verminderde productiviteit meenamen ($n = 22$). Verder zorgden de internet interventies voor mentale stoornissen voor een statistisch significante toename in kwaliteit van leven vergeleken met de controle condities, hoewel het verschil zeer klein was. Ondanks dat er geen statistisch significant verschil in kosten en slechts een verwaarloosbaar verschil in effecten (toename in kwaliteit van leven) werd gevonden, werden internet interventies toch kosteneffectief bevonden in vergelijking met controle condities. Dit kwam doordat de balans tussen de heilzame effecten en kosten bij internet interventies structureel gunstiger bleek dan bij controle condities; ze zorgden voor een toename in kwaliteit van leven tegen een relatief gunstige prijs (of zorgde in sommige gevallen zelfs voor kostenbesparingen). De resultaten geven aan dat e-mental health interventies voordelig lijken vanuit een economisch perspectief en het voordelen kan opleveren om zulke interventies in de maatschappij uit te rollen. Wel maakte de heterogeniteit tussen de geïncludeerde studies het moeilijk om definitieve uitspraken hierover

te doen.

Het was verassend dat in de meta-analyse werd gevonden dat internet interventies maar een minimaal voordeel opleverde in termen van kwaliteit van leven ten opzichte van controle condities. Het kan zijn dat instrumenten om kwaliteit van leven te meten die vaak gebruikt worden in economische evaluaties van interventies niet geschikt zijn voor situaties buiten de somatische zorg, zoals ouderenzorg, chronische ziekten, kwaliteit van de woonomgeving en psychiatrie. De ICECAP-A vragenlijst wordt steeds vaker als aanvulling gebruikt op andere kwaliteit-van-leven instrumenten, omdat het een bredere benadering van welzijn hanteert en dus sensitiever zou kunnen zijn voor veranderingen in kwaliteit van leven in verschillende situaties. De ICECAP-A bevat vijf domeinen – (1) stabiliteit, (2) liefde en ondersteuning, (3) onafhankelijkheid, (4) prestaties en vooruitgang en (5) plezier en genieten – met elk vier niveaus (van 1 'een domein geheel niet kunnen ervaren' tot 4 'een domein volledig kunnen ervaren'). In **hoofdstuk 5** werd gekeken wat de Nederlandse versie van de ICECAP-A precies meet en hoe betrouwbaar dat gebeurt. Op basis van een groep van 1002 deelnemers werd bevonden dat de ICECAP-A tamelijk sterk gerelateerd was aan gebruikelijke kwaliteit-van-leven instrumenten zoals de EQ-5D, maar ook in sommige opzichten ervan verschilde. Zo was de ICECAP-A vooral sterk gerelateerd aan een maat voor zelf-effectiviteit (self-efficacy). Ook kon de ICECAP-A groepen onderscheiden met variërende mate van kwaliteit van leven, wat een belangrijke eigenschap is bij het meten van veranderingen bij mensen die een behandeling hebben gekregen. Samengevat heeft de ICECAP-A betrekking op een concept rakend aan gezondheid, maar meer gerelateerd aan zelf-effectiviteit en welzijn. Daarom zou het andere kwaliteit-van-leven instrumenten kunnen aanvullen bij economische evaluaties, zeker als het gaat om interventies buiten een ziekenhuis context.

In **hoofdstuk 6** werd een tarief ontwikkeld voor de Nederlandse versie van de ICECAP-A, zodat de vragenlijst op betrouwbare wijze gebruikt kan worden bij toekomstige economische evaluaties van interventies in Nederland. Een tarief bevat informatie om ruwe scores te transformeren naar index scores, waarbij belangrijkere domeinen meer gewicht krijgen. Als iemand bijvoorbeeld verbetert op het gebied van plezier en genieten en iemand anders evenveel verbetert op het gebied van onafhankelijkheid zullen zij wellicht niet dezelfde verbetering in algehele kwaliteit van leven ervaren. Door antwoorden van 1002 deelnemers op een zogeheten 'discrete keuze experiment' te analyseren, kon worden gededuceerd dat alle vijf domeinen van de ICECAP-A belangrijk werden bevonden door de algehele Nederlandse bevolking. Stabiliteit – de mate waarin iemand zich op diens plek en veilig voelt – en plezier – de mate waarin iemand plezier kan maken en kan genieten – werden als belangrijkst ervaren. Verder werd gevonden dat het verbeteren van lage niveaus van een domein (bijvoorbeeld van niveau 1 naar 2) als waardevoller werd ervaren dan het verbeteren van gemiddelde of hoge niveaus (bijvoorbeeld van niveau 3 naar 4). Mensen helpen met een lage score op de domeinen van de ICECAP-A, ten opzichte van mensen met een hoge score, zal dus over het geheel meer welzijn opleveren. In vergelijking met het Verenigd Koninkrijk lijkt de Nederlands populatie plezier maken iets belangrijker en prestaties en vooruitgang iets minder belangrijk te vinden. Interventies die plezier maken en genieten kunnen bevorderen zullen wellicht een iets grotere impact op welzijn hebben in Nederland dan in het Verenigd Koninkrijk. Nu er een tarief is ontwikkeld voor de ICECAP-A voor de algemene Nederlandse bevolking kan het

instrument worden gebruikt in economische evaluaties van behandelingen en interventies in Nederland en kan het andere kwaliteit-van-leven instrumenten complementeren.

Na breder te hebben gekeken naar kosteneffectiviteit van e-mental health interventies en mogelijke verbeteringen in het meten van heilzame effecten van interventies door het toevoegen van de ICECAP-A, lag in **hoofdstuk 7** wederom de focus op internet interventies voor eetstoornissen. Dit hoofdstuk bouwde voort op de bevindingen over Featback, ondersteuning door een ervaringsdeskundige en hun combinatie door te kijken naar hun kosteneffectiviteit vergeleken met gebruikelijke zorg (wachtlijst controle conditie). Zowel de EQ-5D als de ICECAP-A werden gebruikt om kwaliteit van leven te meten. Verder werden kosten gerelateerd aan de interventie (bijvoorbeeld personeelskosten), gezondheidszorgkosten en kosten gerelateerd aan ziekteverzuim en verminderde productiviteit door klachten per deelnemer in kaart gebracht. Deelnemers in de vier condities lieten vergelijkbare patronen met betrekking tot kwaliteit van leven zien. Kosten waren het laagst in de Featback conditie, maar verschillen tussen de vier condities waren niet statistisch significant. De totale kosten in de Featback, combinatie, ondersteuning en wachtlijst controle conditie waren gemiddeld respectievelijk €16,741, €23,980, €23,620 en €28,479 per deelnemer over een periode van 14 maanden. Resultaten gaven aan dat het aanbieden van alleen Featback de hoogste kans had van de vier condities om efficiënt te zijn. Dat wil zeggen dat de balans tussen het ervaren van verbeteringen in kwaliteit van leven en kosten het gunstigst werd bevonden voor deelnemers in de Featback conditie. De bevindingen waren vergelijkbaar wanneer de EQ-5D of de ICECAP-A werd gebruikt voor het meten van kwaliteit van leven, wat aangaf dat er in dit onderzoek geen voorkeur was voor één van de twee vragenlijsten. Samengevat tonen de resultaten dat Featback zonder ondersteuning door een ervaringsdeskundige de voorkeur heeft vanuit een (maatschappelijk) economisch perspectief.

Hoofdstuk 8 - Implicaties
Dit hoofdstuk begon met een samenvatting van de onderzoeken die in dit proefschrift zijn beschreven. Daarna werden implicaties en aanbevelingen voor vervolgonderzoek gepresenteerd. Een eerste implicatie betrof de behandelkloof die in hoofdstuk 1 werd besproken. De bevindingen in dit proefschrift bevestigden namelijk de potentie van internet interventies voor eetstoornissen om de grote groep mensen met eetstoornis gerelateerde klachten te bereiken die momenteel nog geen passende zorg ontvangen. Zelfs mensen met langdurige en ernstige klachten kunnen baat hebben bij een laagdrempelige, korte online interventie en vermindering van eetstoornissymptomen ervaren. Een schaalbare interventie, zoals Featback, kan een belangrijke rol spelen in het verkleinen van de behandelkloof voor eetstoornissen. Ook vanuit een economisch perspectief lijkt het verstandig om internet interventies te implementeren. In dit proefschrift werd namelijk gevonden dat ze de kwaliteit van leven bevorderden tegen een relatief goedkope prijs. Een internet interventie voor eetstoornissen zoals Featback breed uitrollen zal daarom niet alleen heilzaam zijn voor de persoon die de interventie ontvangt, maar ook voordelig zijn voor de maatschappij als geheel. Beleidsmakers, clinici en onderzoekers wordt aangeraden om na te denken over het verspreiden van zulke interventies. Daarnaast worden subsidie verstrekkers, patiëntenorganisaties, ggz instellingen, beleidsmakers, gemeentes en zorgverzekeraars aangemoedigd om een functioneel financieel systeem te bewerkstelligen om de implementatie van bewezen effectieve internet interventies te stim-

uleren. Deze innovatieve interventies lijken de investering namelijk waard te zijn of zelfs kosten te kunnen besparen.

Ondersteuning door een ervaringsdeskundige toevoegen aan Featback leek de effectiviteit niet te bevorderen, maar wel de tevredenheid met de interventie. Vergelijkbare resultaten met betrekking tot Featback zijn al eerder gevonden. Featback aanbieden zonder ondersteuning lijkt daarom aangewezen, ook omdat het volledig automatisch werkt en zeer schaalbaar is. Echter, het is belangrijk om te kijken hoe het aanbieden van ondersteuning invloed heeft op het aantal aanmeldingen voor de interventie, voordat definitieve aanbevelingen over implementatie kunnen worden gedaan.

Tot slot gaven de resultaten van dit proefschrift de waarde aan van het doen van economische evaluaties van internet interventies. Het complementeert effectiviteitsonderzoek door een completer beeld te geven van de potentiële voordelen van zulke innovatieve interventies. Onderzoekers en clinici wordt aangeraden kennis te nemen van de toegevoegde waarde en beperkingen van kosteneffectiviteitsonderzoek en, bij het doen van zulk soort onderzoek, vroegtijdig een gezondheidseconoom te betrekken. Een gezondheidseconoom kan ook helpen bij beslissingen over geschikte kwaliteit-van-leven instrumenten (bijvoorbeeld EQ-5D en ICECAP-A). Bevindingen uit dit proefschrift leidden niet tot een voorkeur voor een bepaalde kwaliteit-van-leven vragenlijst bij eetstoornissen.

Hoofdstuk 8 - Sterke punten en beperkingen

Een sterk punt van dit proefschrift was dat de resultaten van het gerandomiseerde onderzoek over Featback en ondersteuning door ervaringsdeskundigen gebaseerd zijn op een groot aantal deelnemers. Participanten werden geworven via online kanalen, zodat de uiteindelijke steekproef representatief was voor toekomstige gebruikers. Echter, generaliseren van de resultaten naar een jongere doelgroep met een kortere ziekteduur is lastig (zie verderop in deze paragraaf). Een ander sterk punt van dit proefschrift was dat de resultaten niet alleen effectiviteit betroffen, maar ook een gedetailleerd beeld schetsten van de kosteneffectiviteit van internet interventies in vergelijking met gebruikelijke zorg. Dit zorgde voor een beter begrip van de waarde van zulke interventies vanuit een economisch perspectief, wat belangrijk is voor beslissingen over de financiering en implementatie.

Verschillende beperkingen dienen ook genoemd te worden. Ten eerste, niet alle deelnemers aan het onderzoek vulden elke vragenlijst in (onderzoek drop-out) of maakten gebruik van alle mogelijke sessies van Featback of ondersteuning door een ervaringsdeskundige (interventie drop-out). Volgens een vooraf bepaalde definitie had een deelnemer een volledige interventie gevolgd wanner deze vijf van de acht monitoringssessies (Featback conditie), vijf van de acht ondersteuningssessies (ondersteuningsconditie) of beide (combinatie conditie) voltooide. In totaal waren er 156 (59%) deelnemers die een volledige interventie volgde: 74 (84%) in de Featback conditie, 48 (55%) in de ondersteuningsconditie en 34 (38%) in de combinatie conditie. Dit geeft aan dat deelnemers waarschijnlijk een drempel ervoeren om volledig gebruik te maken van een interventie, zeker met betrekking tot ondersteuning door een ervaringsdeskundige. Proactief afspraken plannen met een ervaringsdeskundige via een online systeem was wellicht te veel werk of werd snel vergeten. In vergelijking kostte het invullen van de monitoringsvragenlijsten minder moeite, omdat deelnemers hiervoor alleen op een link hoefde te klikken die ze wekelijks via de mail opgestuurd kregen. Naast deze interventie drop-out was er sprake van onderzoek drop-out. Ongeveer één-derde van de

deelnemers viel uit gedurende het 14-maanden durende onderzoek, wat invloed gehad zou kunnen hebben op de resultaten. Echter, missende waarden vanwege deze drop-out zijn op betrouwbare wijze geschat met behulp van multipele imputatie methodes. Bovendien waren de resultaten van sensitiviteitsanalyses met alleen deelnemers die het volledige onderzoek volgde vergelijkbaar met die van de hoofdanalyse. Dit heeft voor vertrouwen in de bevindingen gezorgd, waarbij de invloed van zowel interventie als onderzoek drop-out op de resultaten gering is gebleven.

Een tweede beperking betrof de geworven steekproef van het gerandomiseerde onderzoek over Featback en ondersteuning door ervaringsdeskundigen. Deze bestond voornamelijk uit deelnemers met ernstige symptomen en een lange duur van klachten omtrent eten, gewicht en figuur. De resultaten van het onderzoek zijn daarom moeilijk te generaliseren naar individuen met minder ernstige symptomen, terwijl zij wellicht juist baat zouden kunnen hebben bij zulke korte, laagdrempelige interventies.

Ten derde waren uitkomsten van het gerandomiseerde onderzoek alleen gebaseerd op zelfrapportage. Dit beperkte de diagnostische precisie van het meten van klachten. Bovendien kan deze methode geleid hebben tot verstoring van de resultaten, omdat deelnemers bepaalde zaken anders hebben ingevuld of anders hebben herinnerd dan in werkelijkheid het geval. Zelfrapportage was echter nodig om het laagdrempelige en anonieme karakter van de interventies en de anonimiteit van deelnemer te behouden.

Een vierde beperking was dat de studies van de meta-analyse in sommige opzichten van elkaar afweken, zodat het moeilijk is om de waarde te duiden van het samenvoegen van de studies. Heterogeniteit tussen studies komt vaak voor bij meta-analyses, maar lijkt problematischer bij meta-analyses over data die kosteneffectiviteit betreffen, aangezien er meerdere uitkomstmaten (effecten én kosten) zijn en dus meer bronnen van heterogeniteit. Dit probleem werd, in de meta-analyse beschreven in dit proefschrift, ondervangen door (1) duidelijke inclusiecriteria van studies te hanteren om een bepaalde mate van homogeniteit te waarborgen en (2) analyses uit te voeren op subgroepen van de geïncludeerde studies met gemeenschappelijke eigenschappen. Ondanks dat de resultaten met voorzichtigheid geïnterpreteerd moeten worden, geven ze een beeld over kosteneffectiviteit van internet interventies voor mentale klachten ten opzichte van gebruikelijke zorg die de afzonderlijke studies overstijgt.

Hoofdstuk 8 - Toekomstig onderzoek

Laagdrempelige interventies zoals Featback lijken de behandelkloof voor eetstoornissen te kunnen verkleinen, door mensen te bereiken die de reguliere zorg niet bereikt. Echter, deelnemers aan het gerandomiseerde gecontroleerde onderzoek waren al lang en ernstig ziek, wat aangeeft dat het een uitdaging blijft om individuen aan het begin van hun eetstoornis te bereiken. Mensen met eetstoornisklachten eerder bereiken is van groot belang om de last van de eetstoornis, voor zowel het individu als de maatschappij, te verminderen. Daarom is het belangrijk om naast conventionele manieren van verspreiding, zoals het uitdelen van flyers op scholen, andere benaderingen te verkennen en te onderzoeken hoe die de werving voor internet interventies beïnvloeden. Ten eerste zou het interessant kunnen zijn om samen te werken met experts op het gebied van sales en communicatie, aangezien gezondheidszorgprofessionals en onderzoekers meestal niet deskundig zijn op het gebied van niet-wetenschappelijke communicatie. Ten tweede is het in toename mate relevant om sociale

media, zoals Instagram en TikTok, die inmiddels een essentieel onderdeel van de leefwereld van jongeren zijn, te begrijpen en te gebruiken om mensen met beginnende eetstoornisproblematiek te bereiken. Zelf een social media kanaal beginnen om effectieve interventies te promoten of samenwerken met influencers met een groot netwerk zijn twee voorbeelden van mogelijkheden om individuen met eetstoornissymptomen vroegtijdig te bereiken.

Een andere interessante onderzoeksrichting is het toevoegen van ondersteuning aan internet interventies. In dit proefschrift werd naast een verhoogde tevredenheid geen toegevoegde waarde van ondersteuning door een ervaringsdeskundige gevonden. Op basis van de resultaten is er iets voor te zeggen om menselijke ondersteuning weg te laten bij internet interventies zoals Featback en te focussen op het verbeteren van het geautomatiseerde deel van de interventie. Echter, welke invloed het aanbieden van ondersteuning heeft op het aantal aanmeldingen en wie zich aanmeldt, is het onderzoeken waard.

Een derde richting voor toekomstig onderzoek is het voortzetten van kosteneffectiviteitsonderzoek van hoge kwaliteit. Gezien de schaalbaarheid van veel innovatieve interventies, kan kosteneffectiviteitsonderzoek helpen bij het duiden van het belang van zulke interventies vergeleken met andere behandelopties zoals face-to-face behandeling of gebruikelijke maatschappelijke zorg.

Tot slot is het verder personaliseren van internet interventies voor eetstoornissen een vruchtbare onderzoeksrichting. Technologische toepassingen worden steeds veelzijdiger, waardoor interventies zich beter kunnen aanpassen aan de behoefte van gebruikers. Beschikbare innovaties maken het mogelijk om grote groepen mensen aan te spreken en effectieve zorg te bieden. Bij het opnemen van nieuwe technologieën kan het heilzaam zijn om samen te werken met de beoogde eindgebruikers. Zo'n co-creatie proces betrekt gebruikers vroeg in de ontwikkelingsfase en helpt bij het identificeren van daadwerkelijke behoeftes van de eindgebruikers, wat leidt tot een overtuigend en gebruikersvriendelijk design. Innovaties op het gebied van Virtual Reality, machine learning, kunstmatige intelligentie en gamification zouden kunnen worden ingezet om bestaande interventies te verbeteren. Zo is er een chatbot ontwikkeld die een menselijk chatgesprek simuleert en die effectief werd bevonden in het verminderen van eetstoornissymptomen en zorgen over gewicht en figuur. Een andere benadering die gebruik maakt van smartphone technologie vormt just-in-time (JIT) interventies. Hierbij worden data veelvuldig verzameld, bijvoorbeeld door meerdere dagelijkse korte vragenlijsten over stemming en welzijn of zelfs passief door sensoren op een smartphone. Deze data worden vervolgens gebruikt in (kunstmatig geleerde) algoritmes om de optimale timing en inhoud van (kleine) interventies of oefeningen voor een specifiek individu te voorspellen en aan te bieden. De twee genoemde voorbeelden dienen ter inspiratie van hoe bestaande interventies op creatieve wijze kunnen worden verbeterd, zonder de noodzaak voor nieuwe, grote onderzoeken die vaak jaren duren om af te ronden. De Featback interventie bevat al enige responsiviteit, waarbij het wekelijkse monitoringsbericht afhangt van de invoer van de gebruiker zelf (personalisatie). Echter, de interventie wordt mogelijk effectiever en aantrekkelijker door het toepassen van andere vernieuwingen, zoals een virtuele coach die doelen stelt samen met de gebruiker en de gebruiker motiveert om de tips in de feedback berichten toe te passen, of het belonen van gebruikmaken van de interventie met badges of het vrijspelen van kleding voor een virtuele avatar die gebruikers zien wanneer ze inloggen. Ondersteuning door een ervaringsdeskundige vervangen door een chatbot die altijd beschikbaar is of het sturen van een geschikt feedback bericht op basis van passief verza-

melde data via de smartphone (JIT interventie) zijn andere voorbeelden. Het toepassen van dit soort innovaties zijn waarschijnlijk het meest effectief wanneer ze een wetenschappelijke onderbouwing hebben, gebruikersvriendelijk zijn bevonden in een co-creatie proces en geïmplementeerd worden in samenwerking met (computer) technische experts.

Conclusie

Op basis van de bevindingen in dit proefschrift en eerder onderzoek kan worden gesteld dat laagdrempelige internet interventies voor eetstoornissen, zoals Featback, bestaande behandelopties op drie manieren kunnen complementeren. Ten eerste zijn zulke eHealth interventies herhaaldelijk effectief bevonden in het reduceren van eetstoornissymptomen. Ten tweede kunnen ze individuen met eetstoornisklachten bereiken die momenteel niet worden bereikt door andere vormen van hulp en het zoeken van (professionele) hulp stimuleren. Tot slot is het waarschijnlijk dat internet interventies kosteneffectief zijn vergeleken met gebruikelijke zorg in de maatschappij. Het implementeren van schaalbare en makkelijk toegankelijke interventies zoals Featback zal dan ook helpen bij het verminderen van de last van eetstoornissen, zowel op individueel als maatschappelijk niveau.

References

Aardoom, J. J., Dingemans, A. E., Fokkema, M., Spinhoven, P., & Van Furth, E. F. (2017). Moderators of change in an internet-based intervention for eating disorders with different levels of therapist support: What works for whom? *Behaviour Research and Therapy*, *89*, 66–74. https://doi.org/10.1016/j.brat.2016.11.012

Aardoom, J. J., Dingemans, A. E., Spinhoven, P., & Van Furth, E. F. (2013a). An internet-based intervention for eating disorders consisting of automated computer-tailored feedback with or without supplemented frequent or infrequent support from a coach: Study protocol for a randomized controlled trial. *Trials*, *14*, 340. https://doi.org/10.1186/1745-6215-14-340

Aardoom, J. J., Dingemans, A. E., Spinhoven, P., & Van Furth, E. F. (2013b). Treating eating disorders over the internet: A systematic review and future research directions. *International Journal of Eating Disorders*, *46*(6), 539–552. https://doi.org/10.1002/eat.22135

Aardoom, J. J., Dingemans, A. E., Spinhoven, P., van Ginkel, J. R., de Rooij, M., & van Furth, E. F. (2016). Web-based fully automated self-help with different levels of therapist support for individuals with eating disorder symptoms: A randomized controlled trial. *Journal of Medical Internet Research*, *18*(6), e159. https://doi.org/10.2196/jmir.5709

Aardoom, J. J., Dingemans, A. E., & Van Furth, E. F. (2016). E-health interventions for eating disorders: Emerging findings, issues, and opportunities. *Current Psychiatry Reports*, *18*(4), 1–8. https://doi.org/10.1007/s11920-016-0673-6

Aardoom, J. J., Dingemans, A. E., van Ginkel, J. R., Spinhoven, P., Van Furth, E. F., & Van den Akker-van Marle, M. E. (2016). Cost-utility of an internet-based intervention with or without therapist support in comparison with a waiting list for individuals with eating disorder symptoms: A randomized controlled trial. *International Journal of Eating Disorders*, *49*(12), 1068–1076. https://doi.org/10.1002/eat.22587

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet research*, *22*(7), e16021. https://doi.org/10.2196/16021

Ådnanes, M., Cresswell-Smith, J., Melby, L., Westerlund, H., Šprah, L., Sfetcu, R., Straßmayr, C., & Donisi, V. (2020). Discharge planning, self-management, and commu-

nity support: Strategies to avoid psychiatric rehospitalisation from a service user perspective. *Patient Education and Counseling*, *103*(5), 1033–1040. https://doi.org/10.1016/j.pec.2019.12.002

Afentou, N., & Kinghorn, P. (2020). A systematic review of the feasibility and psychometric properties of the ICEpop CAPability Measure for Adults and its use so far in economic evaluation. *Value in Health*, *23*(4), 515–526. https://doi.org/10.1016/j.jval.2019.12.010

Ágh, T., Kovács, G., Supina, D., Pawaskar, M., Herman, B. K., Vokó, Z., & Sheehan, D. V. (2016). A systematic review of the health-related quality of life and economic burdens of anorexia nervosa, bulimia nervosa, and binge eating disorder. *Eating and Weight Disorders*, *21*(3), 353–364. https://doi.org/10.1007/s40519-016-0264-x

Ahern, E., Kinsella, S., & Semkovska, M. (2018). Clinical efficacy and economic evaluation of online cognitive behavioral therapy for major depressive disorder: A systematic review and meta-analysis. *Expert Review of Pharmacoeconomics & Outcomes Research*, *18*(1), 25–41. https://doi.org/10.1080/14737167.2018.1407245

Akers, L., Rohde, P., Shaw, H., & Stice, E. (2021). Cost-effectiveness comparison of delivery modalities for a dissonance-based eating disorder prevention program over 4-year follow-up. *Prevention Science*, *22*, 1086–1095. https://doi.org/10.1007/s11121-021-01264-1

Al-Janabi, H., Flynn, T. N., & Coast, J. (2012). Development of a self-report measure of capability wellbeing for adults: The ICECAP-A. *Quality of Life Research*, *21*(1), 167–176. https://doi.org/10.1007/s11136-011-9927-2

Al-Janabi, H., Flynn, T. N., Peters, T. J., Bryan, S., & Coast, J. (2015). Test-retest reliability of capability measurement in the UK general population. *Health Economics*, *24*(5), 625–630. https://doi.org/10.1002/hec.3100

Al-Janabi, H., Peters, T. J., Brazier, J., Bryan, S., Flynn, T. N., Clemens, S., Moody, A., & Coast, J. (2013). An investigation of the construct validity of the ICECAP-A capability measure. *Quality of Life Research*, *22*(7), 1831–1840. https://doi.org/10.1007/s11136-012-0293-5

Alpay, L., Doms, R., & Bijwaard, H. (2019). Embedding persuasive design for self-health management systems in Dutch healthcare informatics education: Application of a theory-based method. *Health Informatics Journal*, *25*(4), 1631–1646. https://doi.org/10.1177/1460458218796642

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)* https://doi.org/10.1176/appi.books.9780890425596

Andersson, G., Carlbring, P., Titov, N., & Lindefors, N. (2019). Internet interventions for adults with anxiety and mood disorders: A narrative umbrella review of recent meta-analyses. *The Canadian Journal of Psychiatry*, *64*(7), 465–470. https://doi.org/10.1177/0706743719839381

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & 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*(3), 288–295. https://doi.org/10.1002/wps.20151

Andrés-Pepiñá, S., Plana, M. T., Flamarique, I., Romero, S., Borràs, R., Julià, L., Gárriz, M., & Castro-Fornieles, J. (2020). Long-term outcome and psychiatric comorbidity of

adolescent-onset anorexia nervosa. *Clinical Child Psychology and Psychiatry*, *25*(1), 33–44. https://doi.org/10.1177/1359104519827629

Arcelus, J., Mitchell, A. J., Wales, J., & Nielsen, S. (2011). Mortality rates in patients with anorexia nervosa and other eating disorders: A meta-analysis of 36 studies. *Archives of General Psychiatry*, *68*(7), 724–731. https://doi.org/10.1001/archgenpsychiatry.2011.74

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*, 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Austin, A., Flynn, M., Richards, K., Hodsoll, J., Duarte, T. A., Robinson, P., Kelly, J., & Schmidt, U. (2021). Duration of untreated eating disorder and relationship to outcomes: A systematic review of the literature. *European Eating Disorders Review*, *29*(3), 329–345. https://doi.org/10.1002/erv.2745

Bachner-Melman, R., Zontag-Oren, E., Zohar, A. H., & Sher, H. (2018). Lives on the line: The online lives of girls and women with and without a lifetime eating disorder diagnosis. *Frontiers in Psychology*, *9*, 2128. https://doi.org/10.3389/fpsyg.2018.02128

Bagepally, B. S., Chaikledkaew, U., Gurav, Y. K., Anothaisintawee, T., Youngkong, S., Chaiyakunapruk, N., McEvoy, M., Attia, J., & Thakkinstian, A. (2020). Glucagon-like peptide 1 agonists for treatment of patients with type 2 diabetes who fail metformin monotherapy: Systematic review and meta-analysis of economic evaluation studies. *BMJ Open Diabetes Research & Care*, *8*, e001020. https://doi.org/10.1136/bmjdrc-2019-001020

Bagepally, B. S., Gurav, Y. K., Anothaisintawee, T., Youngkong, S., Chaikledkaew, U., & Thakkinstian, A. (2019). Cost utility of sodium-glucose cotransporter 2 inhibitors in the treatment of metformin monotherapy failed type 2 diabetes patients: A systematic review and meta-analysis. *Value in Health*, *22*(12), 1458–1469. https://doi.org/10.1016/j.jval.2019.09.2750

Baji, P., Farkas, M., Dobos, Á., Zrubka, Z., Gulácsi, L., Brodszky, V., Rencz, F., & Péntek, M. (2020). Capability of well-being: Validation of the hungarian version of the ICECAP-A and ICECAP-O questionnaires and population normative data. *Quality of Life Research*, *29*, 2863–2874. https://doi.org/10.1007/s11136-020-02542-1

Bandura, A. (1988). Organizational application of social cognitive theory. *Australian Journal of Management*, *13*(2), 275–302. https://doi.org/10.1177/031289628801300210

Barakat, S., Maguire, S., Smith, K. E., Mason, T. B., Crosby, R. D., & Touyz, S. (2019). Evaluating the role of digital intervention design in treatment outcomes and adherence to etherapy programs for eating disorders: A systematic review and meta-analysis. *International Journal of Eating Disorders*, *52*(10), 1077–1094. https://doi.org/10.1002/eat.23131

Basu, A., & Maciejewski, M. L. (2019). Choosing a time horizon in cost and cost-effectiveness analyses. *JAMA*, *321*(11), 1096–1097. https://doi.org/10.1001/jama.2019.1153

Bauer, S., Bilić, S., Ozer, F., & Moessner, M. (2019). Dissemination of an internet-based program for the prevention and early intervention in eating disorders. *Zeitschrift für Kinder-und Jugendpsychiatrie und Psychotherapie*, *48*, 25–32. https://doi.org/10.1024/1422-4917/a000662

Bauer, S., & Goldschmidt, A. B. (2019). Introduction to the special issue on advancing assessment of, and interventions for, eating disorders via innovative uses of technology. *International Journal of Eating Disorders*, *52*, 1073–1076. https://doi.org/10.1002/eat.23164

Bauer, S., Moessner, M., Wolf, M., Haug, S., & Kordy, H. (2009). Es(s)prit–an internet-based programme for the prevention and early intervention of eating disorders in college students. *British Journal of Guidance & Counselling*, *37*(3), 327–336. https://doi.org/10.1080/03069880902957049

Bauer, S., Winn, S., Schmidt, U., & Kordy, H. (2005). Construction, scoring and validation of the Short Evaluation of Eating Disorders (SEED). *European Eating Disorders Review*, *13*, 191–200. https://doi.org/10.1002/erv.637

Baumeister, H., Reichler, L., Munzinger, M., & Lin, J. (2014). The impact of guidance on internet-based mental health interventions: A systematic review. *Internet Interventions*, *1*(4), 205–215. https://doi.org/10.1016/j.invent.2014.08.003

Becker, A. E., Hadley, A. A., Perloe, A., Fay, K., & Striegel-Moore, R. H. (2010). A qualitative study of perceived social barriers to care for eating disorders: Perspectives from ethnically diverse health care consumers. *International Journal of Eating Disorders*, *43*(7), 633–647. https://doi.org/10.1002/eat.20755

Beintner, I., Jacobi, C., & Taylor, C. B. (2012). Effects of an internet-based prevention programme for eating disorders in the USA and Germany: A meta-analytic review. *European Eating Disorders Review*, *20*(1), 1–8. https://doi.org/10.1002/erv.1130

Bennett, G. G., & Glasgow, R. E. (2009). The delivery of public health interventions via the internet: Actualizing their potential. *Annual Review of Public Health*, *30*, 273–292. https://doi.org/10.1146/annurev.publhealth.031308.100235

Berg, K. C., Peterson, C. B., Frazier, P., & Crow, S. J. (2011). Psychometric evaluation of the eating disorder examination and eating disorder examination-questionnaire: A systematic review of the literature. *International Journal of Eating Disorders*, *45*(3), 428–438. https://doi.org/10.1002/eat.20931

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

Beveridge, J., Phillipou, A., Edwards, K., Hobday, A., Hilton, K., Wyett, C., Saw, A., Graham, G., Castle, D., Brennan, L., Harrison, P., de Gier, R., Warren, N., Hanly, F., Torrens-Witherow, B., Newton, R. L., & group committee, S. (2018). Peer mentoring for eating disorders: Evaluation of a pilot program. *Pilot and Feasibility Studies*. https://doi.org/10.1186/s40814-018-0268-6

Bewell, C. V., & Carter, J. C. (2008). Readiness to change mediates the impact of eating disorder symptomatology on treatment outcome in anorexia nervosa. *International Journal of Eating Disorders*, *41*(4), 368–371. https://doi.org/10.1002/eat.20513

Bode, K., Götz von Olenhusen, N. M., Wunsch, E. M., Kliem, S., & Kröger, C. (2017). Population-based cost-offset analyses for disorder-specific treatment of anorexia nervosa and bulimia nervosa in Germany. *International Journal of Eating Disorders*, *50*(3), 239–249. https://doi.org/10.1002/eat.22686

Bogosian, A., Chadwick, P., Windgassen, S., Norton, S., McCrone, P., Mosweu, I., Silber, R., & Moss-Morris, R. (2015). Distress improves after mindfulness training for progressive ms: A pilot randomised trial. *Multiple Sclerosis Journal*, *21*(9), 1184–1194. https://doi.org/10.1177/1352458515576261

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis version 3*. Biostat.

Bouwmans, C., Krol, M., Severens, H., Koopmanschap, M., Brouwer, W., & Hakkaart-van Roijen, L. (2015). The IMTA Productivity Cost Questionnaire: A standardized instrument for measuring and valuing health-related productivity losses. *Value in Health*, *18*, 753–758. https://doi.org/10.1016/j.jval.2015.05.009

Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology*, *51*(11), 1115–1128. https://doi.org/10.1016/s0895-4356(98)00103-6

Bryan, A. E. B., & Arkowitz, H. (2015). Meta-analysis of the effects of peer-administered psychosocial interventions on symptoms of depression. *American Journal of Community Psychology*, *55*, 455–471. https://doi.org/10.1007/s10464-015-9718-y

Buntrock, C., Berking, M., Smit, F., Lehr, D., Nobis, S., Riper, H., Cuijpers, P., & Ebert, D. (2017). Preventing depression in adults with subthreshold depression: Health-economic evaluation alongside a pragmatic randomized controlled trial of a web-based intervention. *Journal of Medical Internet Research*, *19*(1), e5. https://doi.org/10.2196/jmir.6587

Burda, B. U., O'Connor, E. A., Webber, E. M., Redmond, N., & Perdue, L. A. (2017). Estimating data from figures with a web-based program: Considerations for a systematic review. *Research Synthesis Methods*, *8*(3), 258–262. https://doi.org/10.1002/jrsm.1232

Burger, F., Neerincx, M. A., & Brinkman, W. P. (2020). Technological state of the art of electronic mental health interventions for major depressive disorder: Systematic literature review. *Journal of Medical Internet Research*, *22*(1), e12599. https://doi.org/10.2196/12599

Byford, S., & Sefton, T. (2003). Economic evaluation of complex health and social care interventions. *National Institute Economic Review*, *186*(1), 98–108. https://doi.org/10.1177/002795010300100114

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *Journal of Abnormal and Social Psychology*, *62*, 713–715. https://doi.org/10.1037/h0044721

Campos, F. A. L., de Sousa, A. R. P., Rodrigues, V. P. C., Marques, A. J. P. S., Dores, A. A. M. R., & Queirós, C. M. L. (2014). Peer support for people with mental illness. *Revista de Psiquiatria Clínica*, *41*(2), 49–55. https://doi.org/10.1590/0101-60830000000009

Cardi, V., Albano, G., Ambwani, S., Cao, L., Crosby, R. D., Macdonald, P., Schmidt, U., & Treasure, J. (2019). A randomised clinical trial to evaluate the acceptability and efficacy of an early phase, online, guided augmentation of outpatient care for adults with anorexia nervosa. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S0033291719002824

Cardi, V., Ambwani, S., Crosby, R. D., Macdonald, P., Todd, G., Park, J., Moss, S., Schmidt, U., & Treasure, J. (2015). Self-help and recovery guide for eating disorders (shared):

Study protocol for a randomized controlled trial. *Trials*. https://doi.org/10.1186/s13063-015-0701-6

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2017). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: An updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115

Carr-Hill, R. A. (1989). Assumptions of the QALY procedure. *Social Science & Medicine*, *29*(3), 469–477. https://doi.org/10.1016/0277-9536(89)90296-7

Chaiyakittisopon, K., Pattanaprateep, O., Ruenroengbun, N., Sapankaew, T., Ingsathit, A., McKay, G. J., Attia, J., & Thakkinstian, A. (2021). Evaluation of the cost-utility of phosphate binders as a treatment option for hyperphosphatemia in chronic kidney disease patients: A systematic review and meta-analysis of the economic evaluations. *The European Journal of Health Economics*, *22*(4), 571–584. https://doi.org/10.1007/s10198-021-01275-3

Chen, G., Ratcliffe, J., Kaambwa, B., McCaffrey, N., & Richardson, J. (2018). Empirical comparison between capability and two health-related quality of life measures. *Social Indicators Research*, *140*, 175–190. https://doi.org/10.1007/s11205-017-1788-9

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., & Muller, K. E. (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine*, *29*(4), 504–520. https://doi.org/10.1002/sim.3775

Ciuca, A. M., Berger, T., Crișan, L. G., & Miclea, M. (2018). Internet-based treatment for panic disorder: A three-arm randomized controlled trial comparing guided (via real-time video sessions) with unguided self-help treatment and a waitlist control. paxpd study results. *Journal of Anxiety disorders*, *56*, 43–55. https://doi.org/10.1016/j.janxdis.2018.03.009

Coast, J. (2004). Is economic evaluation in touch with society's health values? *British Medical Journal*, *329*(7476), 1233–1236. https://doi.org/10.1136/bmj.329.7476.1233

Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J., & Peters, T. J. (2008). Valuing the ICECAP capability index for older people. *Social Science & Medicine*, *67*(5), 874–882. https://doi.org/10.1016/j.socscimed.2008.05.015

Coast, J., Smith, R. D., & Lorgelly, P. (2008). Welfarism, extra-welfarism and capability: The spread of ideas in health economics. *Social Science & Medicine*, *67*(7), 1190–1198. https://doi.org/10.1016/j.socscimed.2008.06.027

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.

Collins, R. L. (1996). For better or worse: The impact of upward social comparison on self-evaluations. *Psychological Bulletin*, *119*(1), 51–69. https://doi.org/10.1037/0033-2909.119.1.51

Cookson, R. (2005). QALYs and the capability approach. *Health Economics*, *14*(8), 817–829. https://doi.org/10.1002/hec.975

Covey, J. (2014). The role of dispositional factors in moderating message framing effects. *Health Psychology*, *33*(1), 52–65. https://doi.org/10.1037/a0029305

Crespo, C., Monleon, A., Diaz, W., & Ríos, M. (2014). Comparative efficiency research (COMER): Meta-analysis of cost-effectiveness studies. *BMC Medical Research Methodology*, *14*(1), 139. https://doi.org/10.1186/1471-2288-14-139

Crombie, I. K., Irvine, L., Williams, B., Sniehotta, F. F., Petrie, D. J., Jones, C., Norrie, J., Evans, J. M. M., Emslie, C., Rice, P. M., Slane, P. W., Humphris, G., Ricketts, I. ., Melson, A. J., Donnan, P. T., McKenzie, A., Huang, L., & Achison, M. (2018). Text message intervention to reduce frequency of binge drinking among disadvantaged men: The TRAM RCT. *Public Health Research*, *6*(6). https://doi.org/10.3310/phr06060

Crow, S. J., Mitchell, J. E., Crosby, R. D., Swanson, S. A., Wonderlich, S., & Lancaster, K. (2009). The cost effectiveness of cognitive behavioral therapy for bulimia nervosa delivered via telemedicine versus face to face. *Behaviour Research and Therapy*, *47*(6), 451–453. https://doi.org/10.1016/j.brat.2009.02.006

Cuijpers, P. (2016). *Meta-analyses in mental health research: A practical guide*. Vrije Universiteit Amsterdam.

Curry, O., & Dunbar, R. I. M. (2013). Do birds of a feather flock together? the relationship between similarity and altruism in social networks. *Human Nature*, *24*(3), 336–347. https://doi.org/10.1007/s12110-013-9174-z

Dahlander, L., & McFarland, D. A. (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly*, *58*(1), 69–110. https://doi.org/10.1177/0001839212474272

De Boer, M. R., Waterlander, W. E., Kuijper, L. D., Steenhuis, I. H. M., & Twisk, J. W. R. (2015). Testing for baseline differences in randomized controlled trials: An unhealthy research behavior that is hard to eradicate. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(4). https://doi.org/10.1186/s12966-015-0162-z

De la Rie, S., Noordenbos, G., Donker, M., & Van Furth, E. F. (2006). Evaluating the treatment of eating disorders from the patient's perspective. *International Journal of Eating Disorders*, *39*(8), 667–676. https://doi.org/10.1002/eat.20317

Dear, B. F., Zou, J. B., Ali, S., Lorian, C. N., Johnston, L., Sheehan, J., Staples, L. G., Gandy, M., Fogliati, V. J., Klein, B., & Titov, N. (2015). Clinical and cost-effectiveness of therapist-guided internet-delivered cognitive behavior therapy for older adults with symptoms of anxiety: A randomized controlled trial. *Behavior Therapy*, *46*(2), 206–217. https://doi.org/10.1016/j.beth.2014.09.007

Deloitte Access Economics. (2020). The social and economic cost of eating disorders in the united states of america: A report for the strategic training initiative for the prevention of eating disorders and the academy for eating disorders. https://www.hsph.harvard.edu/striped/report-economic-costs-of-eating-disorders/

Deluca, P., Coulton, S., Fasihul Alam, M., Boniface, S., Donoghue, K., Gilvarry, E., Kaner, E., Lynch, E., Maconochie, I., McArdle, P., McGovern, R., Newbury-Birch, D., Patton, R., Pellatt-Higgins, T., Philips, C., Philips, T., Pockett, R., Russell, I. T., Strang, J., & Drummond, C. (2020). Screening and brief interventions for adolescent alcohol use disorders presenting through emergency departments: A research programme including two RCTs. *Programme Grants for Applied Research*, *8*(2). https://doi.org/10.3310/pgfar08020

Dennis, C. (2003). Peer support within a health care context: A concept analysis. *International Journal of Nursing Studies*, *40*, 321–332. https://doi.org/10.1016/S0020-7489(02)00092-5

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dibble, J. L., Levine, T. R., & Park, H. S. (2011). The unidimensional relationship closeness scale (URCS): Reliability and validity evidence for a new measure of relationship closeness. *Psychological Assessment*, *24*(3), 565–72. https://doi.org/10.1037/a0026265

Dirksen, C., & Evers, S. (2016). *Broad consultation as part of the standardization of economic evaluation research in the youth sector*. Maastricht University / Trimbos Instituut.

Dixon, P., Hollinghurst, S., Edwards, L., Thomas, C., Foster, A., Davies, B., Gaunt, D., Montgomery, A. A., & Salisbury, C. (2016). Cost-effectiveness of telehealth for patients with depression: Evidence from the healthlines randomised controlled trial. *BJPsych Open*, *2*(2), 262–269. https://doi.org/10.1192/bjpo.bp.116.002907

Dölemeyer, R., Tietjen, A., Kersting, A., & Wagner, B. (2013). Internet-based interventions for eating disorders in adults: A systematic review. *BMC Psychiatry*, *13*(1), 207. https://doi.org/10.1186/1471-244X-13-207

Donker, T., Blankers, M., Hedman, E., Ljótsson, B., Petrie, K., & Christensen, H. (2015). Economic evaluations of internet interventions for mental health: A systematic review. *Psychological Medicine*, *45*, 3357–3376. https://doi.org/10.1017/S0033291715001427

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. *Advances in Data Analysis and Classification*, *8*, 403–425. https://doi.org/10.1007/s11634-013-0159-x

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics*, *8*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of webplotdigitizer in extracting graphed data. *Behavior Modification*, *41*(2), 323–339. https://doi.org/10.1177/0145445516673998

Duarte, A., Walker, S., Littlewood, E., Brabyn, S., Hewitt, C., Gilbody, S., & Palmer, S. (2017). Cost-effectiveness of computerized cognitive-behavioural therapy for the treatment of depression in primary care: Findings from the randomised evaluation of the effectiveness and acceptability of computerised therapy (REEACT) trial. *Psychological Medicine*, 1–11. https://doi.org/10.1017/S0033291717000289

Ebert, D. D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., Brugnera, A., Øverland, S., Trebbi, G., Jensen, K. L., Kaehlke, F. (b. o. t. E. E.-H. T., Baumeister, H., & Taylor, J. (2018). Internet-and mobile-based psychological interventions: Applications, efficacy, and potential for improving mental health. *European Psychologist*, *23*(2), 167–187. https://doi.org/10.1027/1016-9040/a000318

Eddy, K. T., Tabri, N., Thomas, J. J., Murray, H. B., Keshaviah, A., Hastings, E., Edkins, E., Krishna, M., Herzog, D. B., Keel, P. K., & Franko, D. L. (2017). Recovery from anorexia nervosa and bulimia nervosa at 22-year follow-up. *The Journal of Clinical Psychiatry*, *78*(2), 184–189. https://doi.org/10.4088/JCP.15m10393

Engel, L., Mortimer, D., Bryan, S., Lear, S. A., & Whitehurst, D. G. T. (2017). An investigation of the overlap between the ICECAP-A and five preference-based health-

related quality of life instruments. *PharmacoEconomics*, *35*(7), 741–753. https://doi.org/10.1007/s40273-017-0491-7

Erskine, H. E., Whiteford, H. A., & Pike, K. M. (2016). The global burden of eating disorders. *Current Opinion in Psychiatry*, *29*(6), 346–353. https://doi.org/10.1097/YCO.0000000000000276

European Values Study. (2017). Atlas of european values. https://www.atlasofeuropeanvalues.eu/maptool.html

EuroQol Group. (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, *16*(e), 199–208. https://doi.org/10.1016/0168-8510(90)90421-9

Everaert, J., Koster, E. H. W., Schacht, R., & De Raedt, R. (2010). Evaluatie van de psychometrische eigenschappen van de rosenberg zelfwaardeschaal in een poliklinisch psychiatrische populatie. *Gedragstherapie*, *43*, 307–317.

Evers, S., Goossens, M., De Vet, H., Van Tulder, M., & Ament, A. (2005). Criteria list for assessment of methodological quality of economic evaluations: Consensus on health economic criteria. *International Journal of Technology Assessment in Health Care*, *21*(2), 240–245. https://doi.org/10.1017/S0266462305050324

Fahkri, M. A. B., Juni, M. H., Rosliza, A. M., & Faisal, I. (2017). Societal perspective in economic evaluation. *International Journal of Public Health and Clinical Sciences*, *4*(4), 41–50.

Fairburn, C. G., & Beglin, S. J. (2008). Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders*, *16*(4), 363–370. https://doi.org/10.1002/1098-108X(199412)16:4⟨363::AID-EAT2260160405⟩3.0.CO;2-%23

Feng, Y. S., Kohlmann, T., Janssen, M. F., & Buchholz, I. (2021). Psychometric properties of the EQ-5D-5L: A systematic review of the literature. *Quality of Life Research*, *30*(3), 647–673. https://doi.org/10.1007/s11136-020-02688-y

Ferwerda, M., Van Beugen, S., Van Middendorp, H., Visser, H., Vonkeman, H., Creemers, M., Van Riel, P., Kievit, W., & Evers, A. (2018). Tailored, therapist-guided internet-based cognitive behavioral therapy compared to care as usual for patients with rheumatoid arthritis: Economic evaluation of a randomized controlled trial. *Journal of Medical Internet Research*, *20*(10), e260. https://doi.org/10.2196/jmir.9997

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

Fleming, T. M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., Lucassen, M., Lau, H. M., Pollmuller, B., & Hetrick, S. (2017). Serious games and gamification for mental health: Current status and promising directions. *Frontiers in Psychiatry*, *7*, 215. https://doi.org/10.3389/fpsyt.2016.00215

Flynn, T. N., Huynh, E., Peters, T. J., Al-Janabi, H., Clemens, S., Moody, A., & Coast, J. (2015). Scoring the ICECAP-A capability instrument: Estimation of a uk general population tariff. *Health Economics*, *24*(3), 258–269. https://doi.org/10.1002/hec.3014

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2007). Best–worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, *26*(1), 171–189. https://doi.org/10.1016/j.jhealeco.2006.04.002

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2010). Using discrete choice experiments to understand preferences for quality of life. *Social Science & Medicine*, *70*, 1957–1965. https://doi.org/10.1016/j.socscimed.2010.03.008

Fogarty, S., Ramjan, L., & Hay, P. (2016). A systematic review and meta-synthesis of the effects and experience of mentoring in eating disorders and disordered eating. *Eating Behaviors*, *21*, 66–75. https://doi.org/10.1016/j.eatbeh.2015.12.004

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*(5), 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Foster, G., Taylor, S. J., Eldridge, S. E., Ramsay, J., & Griffiths, C. J. (2007). Self-management education programmes by lay leaders for people with chronic conditions. *The Cochrane Database of Systematic Reviews*, (4), CD005108. https://doi.org/10.1002/14651858.CD005108.pub2

Franck, E., De Raedt, R., Barbez, C., & Rosseel, Y. (2008). Psychometric properties of the Dutch Rosenberg Self-Esteem Scale. *Psychologica Belgica*, *48*, 25–35. https://doi.org/10.5334/pb-48-1-25

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: A comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS one*, *10*(6), e0129478. https://doi.org/10.1371/journal.pone.0129478

Galmiche, M., Déchelotte, P., Lambert, G., & Tavolacci, M. P. (2019). Prevalence of eating disorders over the 2000–2018 period: A systematic literature review. *The American Journal of Clinical Nutrition*, *109*(5), 1402–1413. https://doi.org/10.1093/ajcn/nqy342

Gärtner, F. R., De Miranda, E., Rijnders, M. E., Freeman, L. M., Middeldorp, J. M., Bloemenkamp, K. W. M., Stiggelbout, A. M., & Van den Akker-Van Marle, M. E. (2015). Good reliability and validity for a new utility instrument measuring the birth experience, the labor and delivery index. *Journal of Clinical Epidemiology*, *68*(10), 1184–1194. https://doi.org/10.1016/j.jclinepi.2015.05.020

Genders, R., & Tchanturia, K. (2010). Cognitive remediation therapy (CRT) for anorexia in group format: A pilot study. *Eating and Weight Disorders*, *15*(4), e234–e239. https://doi.org/10.1007/BF03325304

Geraedts, A. S., Van Dongen, J. M., Kleiboer, A. M., Wiezer, N. M., Van Mechelen, W., Cuijpers, P., & Bosmans, J. E. (2015). Economic evaluation of a web-based guided self-help intervention for employees with depressive symptoms: Results of a randomized controlled trial. *Journal of Occupational and Environmental Medicine*, *57*(6), 666–675. https://doi.org/10.1097/JOM.0000000000000423

Gerhards, S. A. H., De Graaf, L. E., Jacobs, L. E., Severens, J. L., Huibers, M. J. H., Arntz, A., Riper, H., Widdershoven, G., Metsemakers, J. F. M., & Evers, S. M. A. A. (2010). Economic evaluation of online computerised cognitive–behavioural therapy without support for depression in primary care: Randomised trial. *The British Journal of Psychiatry*, *196*(4), 310–318. https://doi.org/10.1192/bjp.bp.109.065748

Goranitis, I., Coast, J., Day, E., Copello, A., Freemantle, N., & Frew, E. (2017). Maximizing health or sufficient capability in economic evaluation? A methodological experiment of treatment for drug addiction. *Medical Decision Making*, *37*(5), 498–511. https://doi.org/10.1177/0272989X16678844

Goranitis, I., Coast, J., Day, E., Copello, A., Freemantle, N., Seddon, J., Bennett, C., & Frew, E. (2016). Measuring health and broader well-being benefits in the context of opiate dependence: The psychometric performance of the ICECAP-A and the EQ-5D-5L. *Value in Health*, *19*(6), 820–828. https://doi.org/10.1016/j.jval.2016.04.010

Griffiths, C. J., Motlib, J., Azad, A., Ramsay, J., Eldridge, S., Feder, G., Khanam, R., Munni, R., Garrett, M., Turner, A., & Barlow, J. (2005). Randomised controlled trial of a lay-led self-management programme for bangladeshi patients with chronic disease. *British Journal of General Practice*, *55*(520), 831–837. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1570787

Griffiths, F., Lindenmeyer, A., Powell, J., Lowe, P., & Thorogood, M. (2006). Why are health care interventions delivered over the internet? A systematic review of the published literature. *Journal of Medical Internet Research*, *8*(2), e10. https://doi.org/10.2196/jmir.8.2.e10

Griffiths, S., Mond, J. M., Li, Z., Gunatilake, S., Murray, S. B., Sheffield, J., & Touyz, S. (2015). Self-stigma of seeking treatment and being male predict an increased likelihood of having an undiagnosed eating disorder. *International Journal of Eating Disorders*, *48*(6), 775–778. https://doi.org/10.1002/eat.22413

Griffiths, S., Rossell, S. L., Mitchison, D., Murray, S. B., & Mond, J. M. (2018). Pathways into treatment for eating disorders: A quantitative examination of treatment barriers and treatment attitudes. *Eating Disorders*, *26*(6), 556–574. https://doi.org/10.1080/10640266.2018.1518086

Groothuis-Oudshoorn, C. G. M., Flynn, T. N., Yoo, H. I., Magidson, J., & Oppe, M. (2018). Key issues and potential solutions for understanding healthcare preference heterogeneity free from patient-level scale confounds. *Patient*, *11*(5), 463–466. https://doi.org/10.1007/s40271-018-0309-5

Haider, S., Chaikledkaew, U., Thavorncharoensap, M., Youngkong, S., Islam, M. A., & Thakkinstian, A. (2019). Systematic review and meta-analysis of cost-effectiveness of rotavirus vaccine in low-income and lower-middle-income countries. *Open Forum Infectious Diseases*, *6*(4). https://doi.org/10.1093/ofid/ofz117

Hakkaart-van Roijen, L., Van der Linden, N., Bouwmans, C. M. A., Kanters, T., & Tan, S. S. (2015). Kostenhandleiding: Methodologie van kostenonderzoek en referentieprijzen voor economische evaluaties in de gezondheidszorg. *Kostenhandleiding*. https://www.zorginstituutnederland.nl/binaries/zinl/documenten/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg/Richtlijn+voor+het+uitvoeren+van+economische+evaluaties+in+de+gezondheidszorg+%28verdiepingsmodules%29.pdf

Hart, L. M., Granillo, M. T., Jorm, A. F., & Paxton, S. J. (2011). Unmet need for treatment in the eating disorders: A systematic review of eating disorder specific treatment seeking among community cases. *Clinical Psychology Review*, *31*(5), 727–735. https://doi.org/10.1016/j.cpr.2011.03.004

Hedman, E., Ljótsson, B., & Lindefors, N. (2012). Cognitive behavior therapy via the internet: A systematic review of applications, clinical efficacy and cost-effectiveness. *Expert Review of Pharmacoeconomics & Outcomes Research*, *12*(6), 645–764. https://doi.org/10.1586/ERP.12.67

Helter, T. M., Coast, J., Łaszewska, A., Stamm, T., & Simon, J. (2020). Capability instruments in economic evaluations of health-related interventions: A comparative review of the literature. *Quality of Life Research*, *29*(6), 1433–1464. https://doi.org/10.1007/s11136-019-02393-5

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., & Sterne, J. A. (2011). The cochrane collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

Hiligsmann, M., Reginster, J.-Y., Tosteson, A. N. A., Bukata, S. V., Saag, K. G., Gold, D. T., Halbout, P., Jiwa, F., Lewiecki, E. M., Pinto, D., Adachi, J. D., Al-Daghri, N., Bruyère, O., Chandran, M., Cooper, C., Harvey, N. C., Einhorn, T. A., Kanis, J. A., Kendler, D. L., … Silverman, S. (2019). Recommendations for the conduct of economic evaluations in osteoporosis: Outcomes of an experts' consensus meeting organized by the European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis and Musculoskeletal Diseases (ESCEO) and the US branch of the International Osteoporosis Foundation. *Osteoporosis International*, *30*(1), 45–57. https://doi.org/10.1007/s00198-018-4744-x

Himmler, S., Soekhai, V., Van Exel, J., & Brouwer, W. (2021). What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *Journal of Choice Modelling*, *38*, 100265. https://doi.org/10.1016/j.jocm.2020.100265

Hollinghurst, S., Peters, T. J., Kaur, S., Wiles, N., Lewisand, G., & Kessler, D. (2010). Cost-effectiveness of therapist-delivered online cognitive–behavioural therapy for depression: Randomised controlled trial. *The British Journal of Psychiatry*, *197*(4), 297–304. https://doi.org/10.1192/bjp.bp.109.073080

Hollingworth, W., McKell-Redwood, D., Hampson, L., & Metcalfe, C. (2013). Cost–utility analysis conducted alongside randomized controlled trials: Are economic end points considered in sample size calculations and does it matter? *Clinical Trials*, *10*(1), 43–53. https://doi.org/10.1177/1740774512465358

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://doi.org/10.2307/4615733

Holst, A., Björkelund, C., Metsini, A., Madsen, J. H., Hange, D., Petersson, E. L. L., Eriksson, M. C. M., Kivi, M., Andersson, P. A., & Svensson, M. (2018). Cost-effectiveness analysis of internet-mediated cognitive behavioural therapy for depression in the primary care setting: Results based on a controlled trial. *BMJ Open*, *8*(6), e019716. https://doi.org/10.1136/bmjopen-2017-019716

Holst-Kristensen, A. W., Fonager, K., & Pedersen, K. M. (2020). Test–retest reliability of ICECAP-A in the adult Danish population. *Quality of Life Research*, *29*(2), 547–557. https://doi.org/10.1007/s11136-019-02331-5

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partitioning in r. *Journal of Machine Learning Research*, *16*, 3905–3909.

Hughes, J., Wood, E., & Smith, G. (2009). Exploring kidney patients' experiences of receiving individual peer support. *Health Expectations*, *12*(4), 396–406. https://doi.org/10.1111/j.1369-7625.2009.00568.x

Hunter, R., Wallace, P., Struzzo, P., Della Vedova, R., Scafuri, F., Tersar, C., Lygidakis, C., McGregor, R., Scafato, E., & Freemantle, N. (2017). Randomised controlled non-inferiority trial of primary care-based facilitated access to an alcohol reduction website: Cost-effectiveness analysis. *BMJ Open*, *7*(11), e014577. https://doi.org/10.1136/bmjopen-2016-014577

Iso-Ahola, S. E., & Mannell, R. C. (2004). Leisure and health. In J. Haworth & A. J. Veal (Eds.), *Work and leisure* (pp. 184–199). Routledge.

Jacobi, C., Abascal, L., & Taylor, C. B. (2004). Screening for eating disorders and high-risk behavior: Caution. *International Journal of Eating Disorders*, *36*(3), 280–295. https://doi.org/10.1002/eat.20048

Jacobson, N. S., & Truax, P. (1992). Clinical significance : A statistical approach to defining meaningful change in psychotherapy research. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 631–648). American Psychological Association. https://doi.org/10.1037/10109-042

Janssen, M. F., Pickard, A. S., Golicky, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., & Busschbach, J. (2012). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: A multi-country study. *Quality of Life Research*, *22*(7), 1717–1727. https://doi.org/10.1007/s11136-012-0322-4

Joesch, J. M., Sherbourne, C. D., Sullivan, G., Stein, M. B., Craske, M. G., & Roy-Byrne, P. (2012). Incremental benefits and cost of coordinated anxiety learning and management for anxiety treatment in primary care. *Psychological Medicine*, *42*(9), 1937. https://doi.org/10.1017/S0033291711002893

Jolstedt, M., Wahlund, T., Lenhard, F., Ljótsson, B., Mataix-Cols, D., Nord, M., Öst, L., Högström, J., Serlachius, E., & Vigerland, S. (2018). Efficacy and cost-effectiveness of therapist-guided internet cognitive behavioural therapy for paediatric anxiety disorders: A single-centre, single-blind, randomised controlled trial. *The Lancet Child & Adolescent Health*, *2*(11), 792–801. https://doi.org/10.1016/S2352-4642(18)30275-X

Juarascio, A. S., Parker, M. N., Lagacey, M. A., & Godfrey, K. M. (2018). Just-in-time adaptive interventions: A novel approach for enhancing skill utilization and acquisition in cognitive behavioral therapy for eating disorders. *International Journal of Eating Disorders*, *51*(8), 826–830. https://doi.org/10.1002/eat.22924

Kählke, F., Buntrock, C., Smit, F., Berking, M., Lehr, D., Heber, E., Funk, B., Riper, H., & Ebert, D. D. (2019). Economic evaluation of an internet-based stress management intervention alongside a randomized controlled trial. *JMIR Mental Health*, *6*(5), e10866. https://doi.org/10.2196/10866

Kanters, T., Bouwmans, C. M. A., Van der Linden, N., Tan, S. S., & Hakkaart-van Roijen, L. (2017). Update of the Dutch manual for costing studies in health care. *PloS ONE*, *12*(11), e0187477. https://doi.org/10.1371/journal.pone.0187477

Kass, A. E., Balantekin, K. N., Fitzsimmons-Craft, E. E., Jacobi, C., Wilfley, D. E., & Taylor, C. B. (2017). The economic case for digital interventions for eating disorders among

united states college students. *International Journal of Eating Disorders*, *50*(3), 250–258. https://doi.org/10.1002/eat.22680

Kawai, K., Kawai, A. T., Wollan, P., & Yawn, B. P. (2017). Adverse impacts of chronic pain on health-related quality of life, work productivity, depression and anxiety in a community-based study. *Family Practice*, *34*(6), 656–661. https://doi.org/10.1093/fampra/cmx034

Kazdin, A. E., Fitzsimmons-Craft, E. E., & Wilfley, D. E. (2017). Addressing critical gaps in the treatment of eating disorders. *International Journal of Eating Disorders*, *50*(3), 170–189. https://doi.org/10.1002/eat.22670

Keel, P. K., & Brown, T. A. (2010). Update on course and outcome in eating disorders. *International Journal of Eating Disorders*, *43*(3), 195–204. https://doi.org/10.1002/eat.20810

Keeley, T., Al-Janabi, H., Lorgelly, P., & Coast, J. (2013). A qualitative assessment of the content validity of the ICECAP-A and EQ-5D-5L and their appropriateness for use in health research. *PloS ONE*, *8*(12), e85287. https://doi.org/10.1371/journal.pone.0085287

Keeley, T., Coast, J., Nicholls, E., Foster, N. E., Jowett, S., & Al-Janabi, H. (2016). An analysis of the complementarity of ICECAP-A and EQ-5D-3L in an adult population of patients with knee pain. *Health and Quality of Life Outcomes*, *14*(1), 1–5. https://doi.org/10.1186/s12955-016-0430-x

Kempen, G. I. J. M., & van Eijk, L. M. (1995). The psychometric properties of the SSL12-I, a short scale for measuring social support in the elderly. *Social Indicators Research*, *35*(3), 303–312. https://doi.org/10.1007/BF01079163

Keski-Rahkonen, A., & Mustelin, L. (2016). Epidemiology of eating disorders in europe: Prevalence, incidence, comorbidity, course, consequences, and risk factors. *Current Opinion in Psychiatry*, *29*(6), 340–345. https://doi.org/10.1097/YCO.0000000000000278

Kew, K. M., Carr, R., & Crossingham, I. (2017). Lay-led and peer support interventions for adolescents with asthma. *The Cochrane database of systematic reviews*, *4*(4), CD012331. https://doi.org/10.1002/14651858.CD012331.pub2

Kiadaliri, A., Eliasson, B., & Gerdtham, U. (2015). Does the choice of EQ-5D tariff matter? A comparison of the Swedish EQ-5D-3L index score with UK, US, Germany and Denmark among type 2 diabetes patients. *Health and Quality of Life Outcomes*, *13*(145), 145. https://doi.org/10.1186/s12955-015-0344-z

Killen, J. D., Taylor, C. B., Hammer, L. D., Litt, I., Wilson, D. M., Rich, T., Hayward, C., Simmonds, B., Kraemer, H., & Varady, A. (1993). An attempt to modify unhealthful eating attitudes and weight regulation practices of young adolescent girls. *International Journal of Eating Disorders*, *13*(4), 369–384. https://doi.org/10.1002/1098-108x(199305)13:4⟨369::aid-eat2260130405⟩3.0.co;2-0

Killen, J. D., Taylor, C. B., Hayward, C., Wilson, D. M., Haydel, K. F., Hammer, L. D., Simmonds, B., Robinsone, T. N., Litt, I., Varady, A., & Kraemer, H. C. (1994). Persuit of thinness and onset of eating disorder symptoms in a community sample of adolescent girls: A three-year prospective analysis. *International Journal of Eating Disorders*, *16*(3), 227–238. https://doi.org/10.1002/1098-108X(199411)16:3⟨227::AID-EAT2260160303⟩3.0.CO;2-L

Kindermann, S., Moessner, M., Bauer, S., Ali, K., & Minarik, C. (2016). Enhancing help-seeking behavior in individuals with eating disorder symptoms via internet: A case report. *Mental Health and Prevention*, *4*(2), 69–74. https://doi.org/10.1016/j.mhp.2016.04.002

Kolovos, S., Kenter, R. M., Bosmans, J. E., Beekman, A. T., Cuijpers, P., Kok, R. N., & Van Straten, A. (2016). Economic evaluation of internet-based problem-solving guided self-help treatment in comparison with enhanced usual care for depressed outpatients waiting for face-to-face treatment: A randomized controlled trial. *Journal of affective disorders*, *200*, 284–292. https://doi.org/10.1016/j.jad.2016.04.025

Kolovos, S., Van Dongen, J. M., Riper, H., Buntrock, C., Cuijpers, P., Ebert, D. D., Geraedts, A. S., Kenter, R. M., Nobis, S., Smith, A., Warmerdam, L., Hayden, J. A., Van Tulder, M. W., & Bosmans, J. E. (2018). Cost effectiveness of guided internet-based interventions for depression in comparison with control conditions: An individual-participant data meta-analysis. *Depression and Anxiety*, *35*, 209–219. https://doi.org/10.1002/da.22714

König, H. H., Bleibler, F., Friederich, H. C., Herpertz, S., Lam, T., Mayr, A., Schmidt, U., Svaldi, J., Zipfel, S., Brettschneider, C., Hilbert, A., De Zwaan, M., & Egger, N. (2018). Economic evaluation of cognitive behavioral therapy and internet-based guided self-help for binge-eating disorder. *International Journal of Eating Disorders*, *51*(2), 155–164. https://doi.org/10.1002/eat.22822

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koopmanschap, M. A., Rutten, F. F., Van Ineveld, B. M., & Van Roijen, L. (1995). The friction cost method for measuring indirect costs of disease. *Journal of Health Economics*, *14*(2), 171–189. https://doi.org/10.1016/0167-6296(94)00044-5

Kraepelien, M., Mattsson, S., Hedman-Lagerlöf, E., Petersson, I. F., Forsell, Y., Lindefors, N., & Kaldo, V. (2018). Cost-effectiveness of internet-based cognitive–behavioural therapy and physical exercise for depression. *BJPsych Open*, *4*(4), 265–273. https://doi.org/10.1192/bjo.2018.38

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, *50*(6), 613–621. https://doi.org/10.1176/appi.psy.50.6.613

Lampard, A. M., Tasca, G. A., Balfour, L., & Bissada, H. (2013). An evaluation of the transdiagnostic cognitive-behavioural model of eating disorders. *European Eating Disorders Review*, *21*(2), 99–107. https://doi.org/10.1002/erv.2214

Latimer, A. E., Brawley, L. R., & Bassett, R. L. (2010). A systematic review of three approaches for constructing physical activity messages: What messages work and what improvements are needed? *International Journal of Behavioral Nutrition and Physical Activity*, *7*, 36. https://doi.org/10.1186/1479-5868-7-36

Launay, J., & Dunbar, R. I. M. (2015). Playing with strangers: Which shared traits attract us most to new people? *PloS ONE*, *10*(6), e0129688. https://doi.org/10.1371/journal.pone.0129688

Lenhard, F., Ssegonja, R., Andersson, E., Feldman, I., Rück, C., Mataix-Cols, D., & Serlachius, E. (2017). Cost-effectiveness of therapist-guided internet-delivered cognitive

behaviour therapy for paediatric obsessive–compulsive disorder: Results from a randomised controlled trial. *BMJ Open*, *7*(5), e015246. https://doi.org/10.1136/bmjopen-2016-015246

Lewis, H. K., & Foye, U. (2022). From prevention to peer support: A systematic review exploring the involvement of lived-experience in eating disorder interventions. *Mental Health Review Journal*, *27*(1), 1–17. https://doi.org/10.1108/MHRJ-04-2021-0033

Linardon, J., Messer, M., Lee, S., & Rosato, J. (2021). Perspectives of e-health interventions for treating and preventing eating disorders: Descriptive study of perceived advantages and barriers, help-seeking intentions, and preferred functionality. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, *26*(4), 1097–1109. https://doi.org/10.1007/s40519-020-01005-3

Linardon, J., Shatte, A., Messer, M., Firth, J., & Fuller-Tyszkiewicz, M. (2020). E-mental health interventions for the treatment and prevention of eating disorders: An updated systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *88*(11), 994–1007. https://doi.org/10.1037/ccp0000575

Lindenberg, K., & Kordy, H. (2015). Wirksamkeit eines gestuften, internetvermittelten ansatzes zur prävention von essstörungen bei schülern der 7. bis 10. klasse. *Kindheit und Entwicklung*, *24*(1), 55–63. https://doi.org/10.1026/0942-5403/a000152

Lindenberg, K., Moessner, M., Harney, J., McLaughlin, O., & Bauer, S. (2011). E-health for individualized prevention of eating disorders. *Clinical Practice and Epidemiology in Mental Health*, *7*, 74–83. https://doi.org/10.2174/1745017901107010074

Lindsäter, E., Axelsson, E., Salomonsson, S., Santoft, F., Ljótsson, B., Åkerstedt, T., Lekander, M., & Hedman-Lagerlöf, E. (2019). Cost-effectiveness of therapist-guided internet-based cognitive behavioral therapy for stress-related disorders: Secondary analysis of a randomized controlled trial. *Journal of Medical Internet Research*, *21*(9), e14675. https://doi.org/10.2196/14675

Linton, M. J., Mitchell, P. M., Al-Janabi, H., Schlander, M., Richardson, J., Iezzi, A., Ubels, J., & Coast, J. (2020). Comparing the German translation of the ICECAP-A capability wellbeing measure to the original English version: Psychometric properties across healthy samples and seven health condition groups. *Applied Research Quality Life*, *15*(3), 651–673. https://doi.org/10.1007/s11482-018-9681-5

Lloyd-Evans, B., Mayo-Wilson, E., Harrison, B., Istead, H., Brown, E., Piling, S., Johnson, S., & Kendall, T. (2014). A systematic review and meta-analysis of randomised controlled trials of peer support for people with severe mental illness. *BMC Psychiatry*, *14*, 39. https://doi.org/10.1186/1471-244X-14-39

Lokman, S., Volker, D., Zijlstra-Vlasveld, M. C., Brouwers, E. P., Boon, B., Beekman, A. T., Smit, F., & Van der Feltz-Cornelis, C. M. (2017). Return-to-work intervention versus usual care for sick-listed employees: Health-economic investment appraisal alongside a cluster randomised trial. *BMJ Open*, *7*(10), e016348. https://doi.org/10.1136/bmjopen-2017-016348

Loucas, C. E., Fairburn, C. G., Whittington, C., Pennant, M. E., Stockton, S., & Kendall, T. (2014). E-therapy in the treatment and prevention of eating disorders: A systematic review and meta-analysis. *Behaviour Research and Therapy*, *63*, 122–131. https://doi.org/10.1016/j.brat.2014.09.011

Lovell, K., Bower, P., Gellatly, J., Byford, S., Bee, P., McMillan, D., Arundel, C., Gilbody, S., Gega, L., Hardy, G., Reynolds, S., Barkham, M., Mottram, P., Lidbetter, N., Pedley, R., Molle, J., Peckham, E., Knopp-Hogger, J., Price, O., ... Roberts, C. (2017). Clinical effectiveness, cost-effectiveness and acceptability of low-intensity interventions in the management of obsessive–compulsive disorder: The obsessive–compulsive treatment efficacy randomised controlled trial (OCTET). *Health Technology Assessment*, *21*(37), 1–132. https://doi.org/10.3310/hta21370

Lundy, J. J., Coons, S. J., Flood, E., & Patel, M. J. (2020). Agreement among paper and electronic modes of the EQ-5D-5L. *The Patient-Patient-Centered Outcomes Research*, *13*(4), 435–443. https://doi.org/10.1007/s40271-020-00419-6

Luszczynska, A., Gutiérrez-Doña, B., & Schwarzer, R. (2005). General self-efficacy in various domains of human functioning: Evidence from five countries. *International Journal of Psychology*, *40*(2), 80–89. https://doi.org/10.1080/00207590444000041

Magidson, J., & Vermunt, J. K. (2007). Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference. *Sawtooth Software*, 139–154.

Mahlke, C. I., Priebe, S. K. F., Heumann, K., Daubmann, A., Wegscheider, K., & Bock, T. (2016). Effectiveness of one-to-one peer support for patients with severe mental illness - a randomised controlled trial. *European Psychiatry*, *42*, 103–110. https://doi.org/10.1016/j.eurpsy.2016.12.007

Marquez, P. V., & Saxena, S. (2016). Making mental health a global priority. In *Cerebrum: The dana forum on brain science*. Dana Foundation.

McCarroll, M. (2012). A pilot study on the effects of mentoring on disorderd eating behavior. *Master's Theses and Doctoral Dissertations*. http://commons.emich.edu/theses/388

McClay, A. C., Waters, L., Schmidt, U., & Williams, C. . B. C. P. 2.-1. (2014). A survey of attitudes towards computerized self-help for eating disorders within a community-based sample. *Behavioural and Cognitive Psychotherapy*, *28*(1), 65–78. https://doi.org/10.1017/S1352465814000484

Mehlsen, M., Hegaard, L., Ørnbøl, E., Jensen, J. S., Fink, P., & Frostholm, L. (2017). The effect of a lay-led, group-based self-management program for patients with chronic pain: A randomized controlled trial of the Danish version of the chronic pain self-management programme. *Pain*, *158*(8), 1437–1445. https://doi.org/10.1097/j.pain.0000000000000931

Melioli, T., Bauer, S., Franko, D. L., Moessner, M., Ozer, F., Chabrol, H., & Rodgers, R. F. (2016). Reducing eating disorder symptoms and risk factors using the internet: A meta-analytic review. *International Journal of Eating Disorders*, *49*(1), 19–31. https://doi.org/10.1002/eat.22477

Mitchell, P. M., Al-Janabi, H., Byford, S., Kuyken, W., Richardson, J., Lezzi, A., & Coast, J. (2017). Assessing the validity of the ICECAP-A capability measure for adults with depression. *BioMed Central*, *17*(46), 1–13. https://doi.org/10.1186/s12888-017-1211-8

Mitchell, P. M., Roberts, T. E., Barton, P. M., & Coast, J. (2015). Assessing sufficient capability: A new approach to economic evaluation. *Social Science & Medicine*, *139*, 71–79. https://doi.org/10.1016/j.socscimed.2015.06.037

Moayeri, F., Dunt, D., Hsueh, Y. S., & Doyle, C. (2019). Cost-utility analysis of telephone-based cognitive behavior therapy in chronic obstructive pulmonary disease (COPD) patients with anxiety and depression comorbidities: An application for willingness to accept concept. *Expert Review of Pharmacoeconomics & Outcomes Research*, *19*(3), 331–340. https://doi.org/10.1080/14737167.2019.1536550

Moessner, M., & Bauer, S. (2017). Maximizing the public health impact of eating disorder services: A simulation study. *International Journal of Eating Disorders*, *50*(12), 1378–1384. https://doi.org/10.1002/eat.22792

Moessner, M., Ozer, F., Bauer, S., & Minarik, C. (2016). Can an internet-based program for the prevention and early intervention in eating disorders facilitate access to conventional professional healthcare? *Journal of Mental Health*, *25*(5), 441–447. https://doi.org/10.3109/09638237.2016.1139064

Moeyaert, M., Maggin, D., & Verkuilen, J. (2016). Reliability, validity, and usability of data extraction programs for single-case research designs. *Behavior Modification*, *40*(6), 874–900. https://doi.org/10.1177/0145445516645763

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Mond, J. M., Hay, P., Rodgers, B., & Owen, C. (2006). Eating disorder examination questionnaire (EDE-Q): Norms for young adult women. *Behaviour Research and Therapy*, *44*(1), 53–62. https://doi.org/10.1016/j.brat.2004.12.003

Montoya, R. M., & Horton, R. S. (2012). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships*, *30*(1), 64–94. https://doi.org/10.1177/0265407512452989

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, *25*, 889–992. https://doi.org/10.1177/0265407508096700

Morriss, R., Patel, S., Malins, S., Guo, B., Higton, F., James, M., Wu, M., Brown, P., Boycott, N., Kaylor-Hughes, C., Morris, M., Rowley, E., Simpson, J., Smart, D., Stubley, M., Kai, J., & Tyrer, H. (2019). Clinical and economic outcomes of remotely delivered cognitive behaviour therapy versus treatment as usual for repeat unscheduled care users with severe health anxiety: A multicentre randomised controlled trial. *BMC Medicine*, *17*(1), 1–13. https://doi.org/10.1186/s12916-019-1253-5

Mühlbacher, A. C., Kaczynski, A., Zweifel, P., & Johnson, F. R. (2016). Experimental measurement of preferences in health and healthcare using best-worst scaling: An overview. *Health Economics Review*, *6*(1), 1–14. https://doi.org/10.1186/s13561-015-0079-x

Murphy, S. M., Campbell, A. N., Ghitza, U. E., Kyle, T. L., Bailey, G. L., Nunes, E. V., & Polsky, D. (2016). Cost-effectiveness of an internet-delivered treatment for substance abuse: Data from a multisite randomized controlled trial. *Drug and Alcohol Dependence*, *161*, 119–126. https://doi.org/10.1016/j.drugalcdep.2016.01.021

Musiat, P., Johnson, C., Atkinson, M., Wilksch, S., & Wade, T. (2022). Impact of guidance on intervention adherence in computerised interventions for mental health problems:

A meta-analysis. *Psychological Medicine*, *52*(2), 229–240. https://doi.org/10.1017/S0033291721004621

Naslund, J. A., Mitchell, L. M., Joshi, U., Nagda, D., & Lu, C. (2022). Economic evaluation and costs of telepsychiatry programmes: A systematic review. *Journal of Telemedicine and Telecare*, *28*(5), 311–330. https://doi.org/10.1177/1357633X20938919

National Institute for Health and Care Excellence. (2016). *The social care guidance manual*.

Nederlandse Zorgautoriteit. (2021). Informatiekaart wachttijden ggz 2021. https://puc.overheid.nl/nza/doc/PUC_648825_22/1/

Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (2016). *Cost-effectiveness in health and medicine. 2nd ed*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190492939.001.0001]ISBN:9780190492939

Nobis, S., Ebert, D., Lehr, D., Smit, F., Buntrock, C., Berking, M., Baumeister, H., Snoek, F., Funk, B., & Riper, H. (2018). Web-based intervention for depressive symptoms in adults with types 1 and 2 diabetes mellitus: A health economic evaluation. *The British Journal of Psychiatry*, *212*(4), 199–206. https://doi.org/10.1192/bjp.2018.10

Noparatayaporn, P., Thavorncharoensap, M., Bagepally, B. S., & Thakkinstian, A. (2021). Incremental net monetary benefit of bariatric surgery: Systematic review and meta-analysis of cost-effectiveness evidences. *Obesity Surgery*, 1–12. https://doi.org/10.1007/s11695-021-05415-9

OECD. (2021a). 2017 PPP benchmark results. https://stats.oecd.org/Index.aspx?DataSetCode=PRICES_CPI#

OECD. (2021b). Consumer price indices (CPIs). https://stats.oecd.org/Index.aspx?DataSetCode=PRICES_CPI#

Organisation for Economic Co-operation and Development. (2014). *Making mental health count: The social and economic costs of neglecting mental health care*. OECD Publishing.

Padín, P. F., González-Rodríguez, R., Verde-Diego, C., & Vázquez-Pérez, R. (2021). Social media and eating disorder psychopathology: A systematic review. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *15*(3). https://doi.org/10.5817/CP2021-3-6

Paganini, S., Teigelkoetter, W., Buntrock, C., & Baumeister, H. (2018). Economic evaluations of internet-and mobile-based interventions for the treatment and prevention of depression: A systematic review. *Journal of Affective Disorders*, *225*, 733–755. https://doi.org/10.1016/j.jad.2017.07.018

Perez, M., Kroon van Diest, A., & Cutts, S. (2014). Preliminary examination of a mentor-based program for eating disorders. *Journal of Eating Disorders*, *2*(1), 24. https://doi.org/10.1186/s40337-014-0024-0

Pfeiffer, P. N., Heisler, M., Piette, J. D., Rogers, M. A. M., & Valenstein, M. (2011). Efficacy of peer support interventions for depression: A meta-analysis. *General Hospital Psychiatry*, *33*(1), 29–36. https://doi.org/10.1016/j.genhosppsych.2010.10.002

Pietersma, S., Van den Akker-Van Marle, M. E., & De Vries, M. (2013). Generic quality of life utility measures in health-care research: Conceptual issues highlighted for the most commonly used utility measures. *International Journal of Wellbeing*, *3*(2), 173–181.

Pittock, A., Hodges, L., & Lawrie, S. M. (2018). The effectiveness of internet-delivered cognitive behavioural therapy for those with bulimic symptoms: A systematic review. *BMC Research Notes*, *11*, 748. https://doi.org/10.1186/s13104-018-3843-2

Potoglou, D., Burge, P., Flynn, T. N., Netten, A., Malley, J., Forder, J., & Brazier, J. E. (2011). Best–worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine*, *72*(10), 1717–1727. https://doi.org/10.1016/j.socscimed.2011.03.027

Powell, J., Williams, V., Atherton, H., Bennett, K., Yang, Y., Davoudianfar, M., Hellsing, A., Martin, A., Mollison, J., Shanyinde, M., Yu, L. M., & Griffiths, K. M. (2020). Effectiveness and cost-effectiveness of a self-guided internet intervention for social anxiety symptoms in a general population sample: Randomized controlled trial. *Journal of Medical Internet Research*, *22*(1), e16804. https://doi.org/10.1177/2055207617702272

R Core Team. (2017). R: A language and environment for statistical computing. www.R-project.org

R Core Team. (2018). R: A language and environment for statistical computing. www.R-project.org

Ramjan, L., Hay, P., & Fogarty, S. (2017). Benefits of a mentoring support program for individials with an eating disorder: A proof of concept pilot program. *BMC Research Notes*, *10*(1), 709. https://doi.org/10.1186/s13104-017-3026-6

Richards, D., Enrique, A., Eilert, N., Franklin, M., Palacios, J., Duffy, D., Earley, C., Chapman, J., Jell, G., Sollesse, S., & Timulak, L. (2020). A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *NPJ Digital Medicine*, *3*(1), 1–10. https://doi.org/10.1038/s41746-020-0293-8

Richardson, G., Kennedy, A., Reeves, D., Bower, P., Lee, V., Middleton, E., Gardner, C., Gately, C., & Rogers, A. (2008). Cost effectiveness of the expert patients programme (EPP) for patients with chronic conditions. *Journal of Epidemiology & Community Health*, *62*(4), 361–367. https://doi.org/10.1136/jech.2006.057430

Riddell, M. A., Dunbar, J. A., Absetz, P., Wolfe, R., Li, H., Brand, M., Aziz, Z., Oldenburg, B., & Investigators, t. A. P. f. P. D. P. (2016). Cardiovascular risk outcome and program evaluation of a cluster randomised controlled trial of a community-based, lay peer led program for people with diabetes. *BioMed Central*, *16*(1), 864. https://doi.org/10.1186/s12889-016-3538-3

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*. https://doi.org/10.1136/bmj.c221

Rogers, A., Kennedy, A., Bower, P., Gardner, C., Gately, C., Lee, V., Reeves, D., & Richardson, G. (2008). The United Kingdom Expert Patients Programme: Results and implications from a national evaluation. *The Medical Journal of Australia*, *189*(10), 21–4. https://doi.org/10.5694/j.1326-5377.2008.tb02205.x

Rohatgi, A. (2020). Webplotdigitizer: Version 4.3. https://automeris.io/WebPlotDigitizer

Röhr, S., Jung, F. U., Pabst, A., Grochtdreis, T., Dams, J., Nagl, M., Renner, A., Hoffmann, R., König, H. H., Kersting, A., & Riedel-Heller, S. G. (2021). A self-help app for

Syrian refugees with posttraumatic stress (Sanadak): Randomized controlled trial. *JMIR mHealth and uHealth*, *9*(1), e24807. https://doi.org/10.2196/24807

Rohrbach, P. J., Dingemans, A. E., Essers, B. A., Furth, E. F., Spinhoven, P., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., & Van den Akker-Van Marle, M. E. (2022). The ICECAP-A instrument for capabilities: Assessment of construct validity and test–retest reliability in a general Dutch population. *Quality of Life Research*, *31*, 687–696. https://doi.org/10.1007/s11136-021-02980-5

Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., Essers, B. A., Spinhoven, P., Furth, E. F., & Van den Akker-Van Marle, M. E. (2021). The ICEpop Capability Measure for Adults instrument for capabilities: Development of a tariff for the Dutch general population. *Value in Health*. https://doi.org/10.1016/j.jval.2021.07.011

Rohrbach, P. J., Dingemans, A. E., Spinhoven, P., Van den Akker-Van Marle, E., Van Ginkel, J. R., Fokkema, M., Moessner, M., Bauer, S., & Van Furth, E. F. (2019). A randomized controlled trial of an internet-based intervention for eating disorders and the added value of expert-patient support: Study protocol. *Trials*, *20*(1), 1–17. https://doi.org/10.1186/s13063-019-3574-2

Rohrbach, P. J., Dingemans, A. E., Van Furth, E. F., Spinhoven, P., Van Ginkel, J. R., Bauer, S., & Van den Akker-Van Marle, M. E. (2022). Cost-effectiveness of three internet-based interventions for eating disorders: Results from a randomized controlled trial. *International Journal of Eating Disorders*, *55*, 1143–1155. https://doi.org/10.1002/eat.23763

Romero-Sanchiz, P., Nogueira-Arjona, R., Garcia-Ruiz, A., Luciano, J., J., G. C., Gili, M., Botella, C., Banos, R., Castro, A., Lopez-Del-Hoyo, Y., Perez Ara, M., Modrego-Alarcon, M., & Mayoral Cleries, F. (2017). Economic evaluation of a guided and unguided internet-based cbt intervention for major depression: Results from a multicenter, three-armed randomized controlled trial conducted in primary care. *PLoS One*, *12*(2), e0172741. https://doi.org/10.1371/journal.pone.0172741

Rosenberg, M. (1965). *Society and the adolescent self-image*. University Press. https://doi.org/10.1515/9781400876136

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, *4*(1), 87–94. https://doi.org/10.1080/07350015.1986.10509497

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696

Samnaliev, M., Noh, H. L., Sonneville, K. R., & Austin, S. B. (2015). The economic burden of eating disorders and related mental health comorbidities: An exploratory analysis using the us medical expenditures panel survey. *Preventive Medicine Reports*, *2*, 32–34. https://doi.org/10.1016/j.pmedr.2014.12.002

Sander, L., Rausch, L., & Baumeister, H. (2016). Effectiveness of internet-based interventions for the prevention of mental disorders: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *3*(3), e38. https://doi.org/10.2196/mental.6061

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schalken, F., Wolters, W., Tilanus, M., Gemert, M., Hoogenhuyze, C., Meijer, E., Kraefft, E., Brenninkmeijer, M., & Postel, M. (2010). *Handbook of online counseling (handboek online hulpverlening)*. Bohn Stafleu van Loghum.

Schlegl, S., Bürger, C., & Voderholzer, U. (2015). The potential of technology-based psychological interventions for anorexia and bulimia nervosa: A systematic review and recommendations for future research. *Journal of Medical Internet Research*, *17*(3). https://doi.org/10.2196/jmir.3554

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology, a user's portfolio: Causal and control beliefs* (pp. 35–37). NFER-NELSON.

Schwarzer, R., & Jerusalem, M. (2014). Everything you wanted to know about general self-efficacy scale but were afraid to ask. *Documentation of the General Self-Efficacy Scale*. http://userpage.fu-berlin.de/~health/faq_gse.pdf

Schwarzer, R., Rochau, U., Saverno, K., Jahn, B., Bornschein, B., Muehlberger, N., Flatscher-Thoeni, M., Schnell-Inderst, P., Sroczynski, G., Lackner, M., Schall, I., Hebborn, A., Pugner, K., Fehervary, A., Brixner, D., & Siebert, U. (2015). Systematic overview of cost–effectiveness thresholds in ten countries across four continents. *Journal of Comparative Effectiveness Research*, *4*(5), 485–504. https://doi.org/10.2217/cer.15.38

Sen, A. (1992). *Inequality reexamined*. Clarendon Press. https://doi.org/10.1093/0198289286.001.0001

Sen, A. (1993). Capability and well-being. In *The quality of life*. Clarendon Press. https://doi.org/10.1093/0198287976.003.0003

Shahtaheri, R. S., Nikfar, S., Sari, A. A., & Nejad, M. S. Y. (2020). Cross-cultural adaptation and psychometric analysis of the Persian version of the ICEpop CAPability measure for Adults capability measure in the Iranian general population. *Value in Health Regional Issues*, *21*, 188–193. https://doi.org/10.1016/j.vhri.2020.01.001

Shields, G. E., & Elvidge, J. (2020). Challenges in synthesising cost-effectiveness estimates. *Systematic Reviews*, *9*(1), 1–7. https://doi.org/10.1186/s13643-020-01536-x

Simoni, J. M., Franks, J. C., Lehavot, K., & Yard, S. S. (2011). Peer interventions to promote health: Conceptual considerations. *American Journal of Orthopsychiatry*, *81*(3), 351–359. https://doi.org/10.1111/j.1939-0025.2011.01103.x

Sinclair, S. J., Blais, M. A., Gansler, D. A., Sandberg, E., Bistis, K., & LoCicero, A. (2010). Psychometric properties of the rosenberg self-esteem scale: Overall and across demographic groups living within the united states. *Evaluation & the Health Professions*, *33*(1), 56–80. https://doi.org/10.1177/0163278709356187

Singh, B., Hawthorne, G., & Vos, T. (2001). The role of economic evaluation in mental health care. *Australian and New Zealand Journal of Psychiatry*, *35*(1), 104–117. https://doi.org/10.1046/j.1440-1614.2001.00845.x

Smink, F. R., Van Hoeken, D., & Hoek, H. W. (2013). Epidemiology, course, and outcome of eating disorders. *Current Opinion in Psychiatry*, *26*(6), 543–548. https://doi.org/10.1097/YCO.0b013e328365a24f

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T.,

McAleenan, A., ... Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Stevens, I. (2001). *Applied multivariate statistics for the social sciences (4th ed.)* Psychology Press. https://doi.org/10.4324/9781410604491

Stice, E., Rohde, P., Shaw, H., & Gau, J. M. (2017). Clinician-led, peer-led, and internet-delivered dissonance-based eating disorder prevention programs: Acute effectiveness of these delivery modalities. *Journal of Consulting and Clinical Psychology*, *85*(9), 883–895. https://doi.org/10.1037/ccp0000211

Stichting 113. (2009). www.113.nl

Tang, C., Xiong, Y., Wu, H., & Xu, J. (2018). Adaptation and assessments of the Chinese version of the ICECAP-A measurement. *Health and Quality of Life Outcomes*, *16*(45), 45. https://doi.org/10.1186/s12955-018-0865-3

Tate, D. F., Finkelstein, E. A., Khavjou, O., & Gustafson, A. (2009). Cost effectiveness of internet interventions: Review and recommendations. *Annals of Behavioral Medicine*, *38*, 40–45. https://doi.org/10.1007/s12160-009-9131-6

Taylor, C. B., Graham, A. K., Flatt, R. E., Waldherr, K., & Fitzsimmons-Craft, E. E. (2021). Current state of scientific evidence on internet-based interventions for the treatment of depression, anxiety, eating disorders and substance abuse: An overview of systematic reviews and meta-analyses. *European Journal of Public Health*, *21*(Supplement1), i3–i10. https://doi.org/10.1093/eurpub/ckz208

Timman, R., Bouwmans, C. M. A., Busschbach, J. J. V., & Hakkaart-van Roijen, L. (2015). Development of the treatment inventory of costs in psychiatric patients: Tic-p mini and midi. *Value in Health*, *18*(8), 994–999. https://doi.org/10.1016/j.jval.2015.07.006

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., Adlam, K., Nordgreen, T., Hogstad Bruvik, K., Hovland, A., Repål, A., Mathiasen, K., Kraepelien, M., Blom, K., Svanborg, C., Lindefors, N., & Kaldo, V. (2018). ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet Interventions*, *13*, 108–115. https://doi.org/10.1016/j.invent.2018.07.006

Titov, N., Dear, B. F., Ali, S., Zou, J. B., Lorian, C. N., Johnston, L., Terides, M. D., Kayrouz, R., Klein, B., Gandy, M., & Fogliati, V. J. (2015). Clinical and cost-effectiveness of therapist-guided internet-delivered cognitive behavior therapy for older adults with symptoms of depression: A randomized controlled trial. *Behavior Therapy*, *46*(2), 193–205. https://doi.org/10.1016/j.beth.2014.09.008

Välimäki, M., Anttila, K., Anttila, M., & Lahti, M. (2017). Web-based interventions supporting adolescents and young people with depressive symptoms: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *5*(12), e180. https://doi.org/10.2196/mhealth.8624

Vall, E., & Wade, T. D. (2015). Predictors of treatment outcome in individuals with eating disorders: A systematic review and meta-analysis. *International Journal of Eating Disorders*, *48*(7), 946–971. https://doi.org/10.1002/eat.22411

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC Press. https://doi.org/10.1201/b11826

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). MICE: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, *49*(1), 78–91. https://doi.org/10.1080/00273171.2013.855890

Van Ginkel, J. R., Linting, M., Rippe, R. C., & Van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Van Hoeken, D., & Hoek, H. W. (2020). Review of the burden of eating disorders: Mortality, disability, costs, quality of life, and family burden. *Current Opinion in Psychiatry*, *33*(6), 521. https://doi.org/10.1097/YCO.0000000000000641

Van Hoof, M., Jeuring, S. F., Jonkers, D. M., Masclee, A. A., Pierik, M. J., Stokroos, R. J., & Joore, M. A. (2016). De Nederlandse vertaling en indruksvalidatie van de ICECAP-A: Meten van kwaliteit van leven volgens de capability-benadering. *Tijdschrift voor gezondheidswetenschappen*, *94*(8), 313–320. https://doi.org/10.1007/s12508-016-0109-z

Van Krugten, F. C. W., Busschbach, J. J. V., Versteegh, M. M., Hakkaart-van Roijen, L., & Brouwer, W. B. F. (2022). The Mental Health Quality of Life Questionnaire (MHQoL): Development and first psychometric evaluation of a new measure to assess quality of life in people with mental health problems. *Quality of Life Research*, *31*, 633–643. https://doi.org/10.1007/s11136-021-02935-w

Van Luenen, S., Kraaij, V., Garnefski, N., Spinhoven, P., & Van den Akker-van Marle, M. E. (2019). Cost-utility of a guided internet-based intervention in comparison with attention only for people with hiv and depressive symptoms: A randomized controlled trial. *Journal of Psychosomatic Research*, *118*, 34–40. https://doi.org/10.1016/j.jpsychores.2019.01.012

Van Son, G. E., Hoek, H. W., Van Hoeken, D., Schellevis, F. G., & Van Furth, E. F. (2012). Eating disorders in the general practice: A case–control study on the utilization of primary care. *European Eating Disorders Review*, *20*(5), 410–413. https://doi.org/10.1002/erv.2185

van Eijk, L. M., Kempen, G. I. J. M., & van Sonderen, F. L. P. (1994). Een korte schaal voor het meten van sociale steun bij ouderen: De SSL-12-I. *Tijdschrift voor Gerontologie en Geriatrie*, *25*(5), 192–196.

Vass, C. M., Wright, S., Burton, M., & Payne, K. (2018). Scale heterogeneity in healthcare discrete choice experiments: A primer. *The Patient-Patient-Centered Outcomes Research*, *11*(2), 167–173. https://doi.org/10.1007/s40271-017-0282-4

Versteegh, M. M., Knies, S., & Brouwer, W. (2016). From good to better: New Dutch guidelines for economic evaluations in healthcare. *Pharmacoeconomics*, *34*(11), 1071–1074. https://doi.org/10.1007/s40273-016-0431-y

Versteegh, M. M., Vermeulen, K. M., Evers, S. M. A. A., de Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health*, *19*(4), 343–352. https://doi.org/10.1016/j.jval.2016.01.003

Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *Lancet Psychiatry*, *3*(2), 171–178. https://doi.org/10.1016/S2215-0366(15)00505-2

Vollert, B., Beintner, I., Musiat, P., Gordon, G., Görlich, D., Nacke, B., Schmidt-Hantke, J., Potterton, R., Spencer, L., Grant, N., Schmidt, U., & Jacobi, C. (2019). Using internet-based self-help to bridge waiting time for face-to-face outpatient treatment for bulimia nervosa, binge eating disorder and related disorders: Study protocol of a randomized controlled trial. *Internet Interventions*, *16*, 26–34. https://doi.org/10.1016/j.invent.2018.02.010

Wade, T. D., & Wilksch, S. M. (2018). Internet eating disorder prevention. *Current Opinion in Psychiatry*, *31*(6), 456–461. https://doi.org/10.1097/YCO.0000000000000450

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*(1), 101–110. https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1⟨101::AID-SIM727⟩3.0.CO;2-E

Warmerdam, L., Smit, F., Van Straten, A., Riper, H., & Cuijpers, P. (2010). Cost-utility and cost-effectiveness of internet-based treatment for adults with depressive symptoms: Randomized trial. *Journal of Medical Internet Research*, *12*(5), e53. https://doi.org/10.2196/jmir.1436

Watson, H. J., McLagan, N., Zerwas, S. C., Crosby, R. D., Levine, M. D., Runfola, C. D., Peat, C. M., Moessner, M., Zimmer, B., Hofmeier, S. M., Hamer, R. M., Marcus, M. D., Bulik, C. M., & Crow, S. J. (2018). Cost-effectiveness of internet-based cognitive-behavioral treatment for bulimia nervosa: Results of a randomized controlled trial. *Journal of Clinical Psychiatry*, *79*(1), 16m11314. https://doi.org/10.4088/JCP.16m11314

Watts, R. D., & Li, I. W. (2019). Use of checklists in reviews of health economic evaluations, 2010 to 2018. *Value in Health*, *22*(3), 377–382. https://doi.org/10.1016/j.jval.2018.10.006

Wayment, H. A., Eiler, B. A., & Cavolo, K. (2020). Self-evaluation strategies in college women trying to lose weight: The relative use of objective and social comparison information. *Frontiers in Psychology*, *11*, 1254. https://doi.org/10.3389/fpsyg.2020.01254

Weissman, R. S., & Rosselli, F. (2017). Reducing the burden of suffering from eating disorders: Unmet treatment needs, cost of illness, and the quest for cost-effectiveness. *Behaviour Research and Therapy*, *88*, 49–64. https://doi.org/10.1016/j.brat.2016.09.006

Whitty, J. A., & Gonçalves, A. S. O. (2018). A systematic review comparing the acceptability, validity and concordance of discrete choice experiments and best–worst scaling for eliciting preferences in healthcare. *The Patient-Patient-Centered Outcomes Research*, *11*(3), 301–317. https://doi.org/10.1007/s40271-017-0288-y

WHOQOL Group. (1995). The World Health Organization quality of life assessment (WHOQOL): Position paper from the World Health Organization. *Social Science & Medicine*, *41*(10), 1403–1409. https://doi.org/10.1016/0277-9536(95)00112-K

World Health Organization. (2021). *Mental health atlas 2020*. https://apps.who.int/iris/bitstream/handle/10665/345946/9789240036703-eng.pdf

Wright, B., Tindall, L., Hargate, R., Allgar, V., Trépel, D., & Ali, S. (2020). Computerised cognitive–behavioural therapy for depression in adolescents: 12-month outcomes of a UK randomised controlled trial pilot study. *BJPsych Open*, *6*(1), e5. https://doi.org/10.1192/bjo.2019.91

Xiong, Y., Wu, H., & Xu, J. (2021). Assessing the reliability and validity of the ICECAP-A instrument in Chinese type 2 diabetes patients. *Health and Quality of Life Outcomes*, *19*(5), 5. https://doi.org/10.1186/s12955-020-01632-5

Yan, C., Rittenbach, K., Souri, S., & Silverstone, P. H. (2019). Cost-effectiveness analysis of a randomized study of depression treatment options in primary care suggests stepped-care treatment may have economic benefits. *BMC Psychiatry*, *19*(1), 240. https://doi.org/10.1186/s12888-019-2223-3

Yang, J., Johnson, F. R., Kilambi, V., & Mohamed, A. F. (2015). Sample size and utility-difference precision in discrete-choice experiments: A meta-simulation approach. *The Journal of Choice Modelling*, *16*(100), 50–57. https://doi.org/10.1016/j.jocm.2015.09.001

Yim, S. H., & Schmidt, U. (2019a). Experiences of computer-based and conventional self-help interventions for eating disorders: A systematic review and meta-synthesis of qualitative research. *International Journal of Eating Disorders*, *52*(10), 1108–1124. https://doi.org/10.1002/eat.23142

Yim, S. H., & Schmidt, U. (2019b). Self-help treatment of eating disorders. *Psychiatric Clinics*, *42*(2), 231–241. https://doi.org/10.1016/j.psc.2019.01.006

Zeiler, M., Kuso, S., Nacke, B., Klesges, L. M., & Waldherr, K. (2021a). Evaluating reach, adoption, implementation and maintenance of internet-based interventions to prevent eating disorders in adolescents: A systematic review. *The European Journal of Public Health*, *31*(Supplement1), i38–i47. https://doi.org/10.1093/eurpub/ckz130

Zeiler, M., Kuso, S., Nacke, B., Klesges, L. M., & Waldherr, K. (2021b). Evaluating reach, adoption, implementation and maintenance of internet-based interventions to prevent eating disorders in adolescents: A systematic review. *The European Journal of Public Health*, *31*(Supplement1), i38–i47. https://doi.org/10.1093/eurpub/ckz130

Zorginstituut Nederland. (2015). *Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg*. https://www.zorginstituutnederland.nl/binaries/zinl/documenten/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg.pdf

Zwaap, J., Knies, S., Van der Meijden, C., Staal, P., & Van der Heiden, L. (2015). Cost-effectiveness in practice. https://english.zorginstituutnederland.nl/binaries/zinl-eng/documents/reports/2015/06/16/cost-effectiveness-in-practice/Cost-effectiveness+in+practice.pdf

Publications

2022

Rohrbach, P. J., Dingemans, A. E., Evers, C. A. J. M., Aardoom, J. J., Van Furth, E. F., Spinhoven, P., Lähde, I., Clemens, F., & Van den Akker-Van Marle, M. E. (in press). Cost-effectiveness of internet interventions compared to treatment as usual for people with mental disorders: a systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*

Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., Essers, B. A. B., Van Furth, E. F., Spinhoven, P., & Van den Akker-van Marle, M. E. (2022). The ICECAP-A instrument for capabilities: assessment of construct validity and test–retest reliability in a general Dutch population. *Quality of Life Research, 31*, 687–696. https://doi.org/10.1007/s11136-021-02980-5

Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G., Van Til, J. A., Essers, B. A., Van Furth, E. F., & Van den Akker-Van, M. E. (2022). The ICEpop Capability Measure for Adults Instrument for Capabilities: Development of a Tariff for the Dutch General Population. *Value in Health, 25*, 125-132. https://doi.org/10.1016/j.jval.2021.07.011

Rohrbach, P. J., Dingemans, A. E., Spinhoven, P., Van Ginkel, J. R., Fokkema, M., Wilderjans, T. F., Bauer, S., & Van Furth, E. F. (2022). Effectiveness of an online self-help program, expert-patient support and their combination for eating disorders: results from a randomized controlled trial. *International Journal of Eating Disorders, 55*, 1361-1373, https://doi.org/10.1002/eat.23785

Rohrbach, P. J., Dingemans, A. E., Van Furth, E. F., Spinhoven, P., Van Ginkel, J. R., Bauer, S., & Van den Akker-Van Marle, M. E. (2022). Cost-effectiveness of three internet-based interventions for eating disorders: results from a randomized controlled trial. *International Journal of Eating Disorders, 55*, 1143-1155. https://doi.org/10.1002/eat.23763

2019

Rohrbach, P. J., Dingemans, A. E., Spinhoven, P., Van den Akker-Van Marle, E., Van Ginkel, J. R., Fokkema, M., Moessner, M., Bauer, S., & Van Furth, E. F. (2019). A randomized controlled trial of an internet-based intervention for eating disorders and the added value of expert-patient support: Study protocol. *Trials, 20*, 1–17. https://doi.org/10.1186/s13063-019-3574-2

Dankwoord

Tijdens dit promotietraject heb ik het genoegen gehad om tijd door te brengen en samen te werken met inspirerende en enthousiaste mensen. Zij hebben deze periode voor mij niet alleen leerzaam, maar ook leuk en bijzonder gemaakt.

Dank je wel Eric van Furth, voor je excellente begeleiding die ervoor heeft gezorgd dat ik met enthousiasme en zelfverzekerdheid onderzoek doe. Ik heb van je geleerd om onderzoek in een breder kader te plaatsen en ben je dankbaar voor onze gesprekken over onderzoek, en natuurlijk muziek, die altijd leidden tot nieuwe energie en ideeën.

Dank je wel Philip Spinhoven, voor je belangrijke rol in mijn promotietraject. Je bent een briljant voorbeeld in hoe onderzoek op een betrouwbare en ethische manier uitgevoerd dient te worden en ik was geraakt door jouw betrokkenheid en de tijd die je voor me had, ondanks alle andere activiteiten in je leven.

Bedankt Alexandra Dingemans, je bent onmisbaar geweest in dit project, niettemin omdat jij me altijd wist te activeren als ik een zetje nodig had. Naast dat ik op professioneel en persoonlijk vlak veel van je heb geleerd, hebben we vooral ook veel plezier gehad, genoten en gelachen.

Mijn dank gaat ook uit naar Elske van den Akker – Van Marle, het was een plezier om met je samen te werken en van je te leren. Je geduldige houding maakt je misschien wel de meest betrouwbare en geruststellende persoon waar ik mee heb samengewerkt.

Veel dank aan alle collega's van GGZ Rivierduinen. Jullie zorgden ervoor dat ik me welkom en gewaardeerd voelde en ik niet bang was om fouten te maken. Mijn waardering gaat vooral uit naar Rita Slof - Op 't Landt, Eline van Bree en Carmen van Hooijdonk. Bedankt voor de fijne tijd en gesprekken die ik met jullie heb gehad.

Zonder de deskundigheid van alle onderzoekers waarmee ik heb samen mogen werken was het proefschrift nooit van dezelfde kwaliteit geweest. In het bijzonder ben ik dankbaar voor de kennis van en fijne samenwerking met Steffi Bauer, Joost van Ginkel, Marjolein Fokkema, Tom Wilderjans, Jiska Aardoom en Catharine Evers.

Een belangrijke bijdrage voor Featback en de chat en e-mail ondersteuning kwam van Masja Ninck Blok en Dominique Pauelsen, Gabriëlle Schmidt, Hiske Visser, Lizzy Steller, Mandy Sleijpen, Marlous van den Beld en Maureen Larde. Ik heb veel bewondering voor jullie betrokkenheid en doorzettingsvermogen.

Het onderzoek is succesvol verlopen mede door de participatie van alle deelnemers. Bedankt voor jullie tijd en openheid.

Ook zonder de genereuze financiële steun van ZonMw (projectnummer 636310001) had dit project niet kunnen plaatsvinden. Dank daarvoor.

Bedankt Daniël en Robin voor jullie ondersteuning als paranimf en voor jullie vriendschap.

Tot slot wil ik mijn vrienden, bandleden, familie, ouders, Herman en Ilse bedanken. Jullie hebben me voortdurend aangemoedigd, gestimuleerd en, wanneer dat nodig was, voor afleiding gezorgd. Bedankt dat jullie een deel van mijn leven zijn.

Curriculum Vitae

Pieter Johannes Rohrbach was born in Alphen aan den Rijn, the Netherlands, on 2 November 1991. After following secondary education at the Groene Hart Lyceum, Alphen aan den Rijn, he started studying Psychology at Leiden University (2009-2012) and completed the master Clinical Psychology (2012-2014). From 2014-2017, Pieter completed his second master in Media Technology at Leiden University cum laude. Meanwhile, he participated in a 3-year minor at the Royal Conservatory of The Hague, practicing jazz drums. In 2017 Pieter started his PhD research at GGZ Rivierduinen Eating Disorders, in collaboration with the Leiden University Medical Center and Leiden University, where he could combine knowledge obtained from the two masters of Clinical Psychology and Media Technology by researching e-mental health for eating disorders. The research described in this doctoral dissertation is a reflection of Pieter's scientific output during his PhD studies.

Chapter A

Supplemental Material for Chapter 2

A.1 SPIRIT Checklist

SPIRIT 2013 Checklist: Recommended items to address in a clinical trial protocol and related documents*

Section/item	Item No	Description	Page
Administrative information			
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym	1
Trial registration	2a	Trial identifier and registry name. If not yet registered, name of intended registry	4
	2b	All items from the World Health Organization Trial Registration Data Set	1
Protocol version	3	Date and version identifier	32
Funding	4	Sources and types of financial, material, and other support	33-34
Roles and responsibilities	5a	Names, affiliations, and roles of protocol contributors	1-2
	5b	Name and contact information for the trial sponsor	1
	5c	Role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities	33-34
	5d	Composition, roles, and responsibilities of the coordinating centre, steering committee, endpoint adjudication committee, data management team, and other individuals or groups overseeing the trial, if applicable (see Item 21a for data monitoring committee)	n.a.
Introduction			
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention	4-10
	6b	Explanation for choice of comparators	4-10
Objectives	7	Specific objectives or hypotheses	8-10

Trial design	8	Description of trial design including type of trial (eg, parallel group, crossover, factorial, single group), allocation ratio, and framework (eg, superiority, equivalence, noninferiority, exploratory)	10-11
Methods: Participants, interventions, and outcomes			
Study setting	9	Description of study settings (eg, community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained	11
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centres and individuals who will perform the interventions (eg, surgeons, psychotherapists)	11
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered	11-13
	11b	Criteria for discontinuing or modifying allocated interventions for a given trial participant (eg, drug dose change in response to harms, participant request, or improving/worsening disease)	n.a.
	11c	Strategies to improve adherence to intervention protocols, and any procedures for monitoring adherence (eg, drug tablet return, laboratory tests)	12-13 and 13-14
	11d	Relevant concomitant care and interventions that are permitted or prohibited during the trial	13
Outcomes	12	Primary, secondary, and other outcomes, including the specific measurement variable (eg, systolic blood pressure), analysis metric (eg, change from baseline, final value, time to event), method of aggregation (eg, median, proportion), and time point for each outcome. Explanation of the clinical relevance of chosen efficacy and harm outcomes is strongly recommended	8-9 and 14-22
Participant timeline	13	Time schedule of enrolment, interventions (including any run-ins and washouts), assessments, and visits for participants. A schematic diagram is highly recommended (see Figure)	10-11 and Figure 1 and Table 1
Sample size	14	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations	25
Recruitment	15	Strategies for achieving adequate participant enrolment to reach target sample size	11
Methods: Assignment of interventions (for controlled trials)			
Allocation:			
Sequence generation	16a	Method of generating the allocation sequence (eg, computer-generated random numbers), and list of any factors for stratification. To reduce predictability of a random sequence, details of any planned restriction (eg, blocking) should be provided in a separate document that is unavailable to those who enrol participants or assign interventions	10-11

Allocation concealment mechanism	16b	Mechanism of implementing the allocation sequence (eg, central telephone; sequentially numbered, opaque, sealed envelopes), describing any steps to conceal the sequence until interventions are assigned	10
Implementation	16c	Who will generate the allocation sequence, who will enrol participants, and who will assign participants to interventions	10
Blinding (masking)	17a	Who will be blinded after assignment to interventions (eg, trial participants, care providers, outcome assessors, data analysts), and how	10
	17b	If blinded, circumstances under which unblinding is permissible, and procedure for revealing a participant's allocated intervention during the trial	n.a.
Methods: Data collection, management, and analysis			
Data collection methods	18a	Plans for assessment and collection of outcome, baseline, and other trial data, including any related processes to promote data quality (eg, duplicate measurements, training of assessors) and a description of study instruments (eg, questionnaires, laboratory tests) along with their reliability and validity, if known. Reference to where data collection forms can be found, if not in the protocol	14-22
	18b	Plans to promote participant retention and complete follow-up, including list of any outcome data to be collected for participants who discontinue or deviate from intervention protocols	22 and 23
Data management	19	Plans for data entry, coding, security, and storage, including any related processes to promote data quality (eg, double data entry; range checks for data values). Reference to where details of data management procedures can be found, if not in the protocol	30-31
Statistical methods	20a	Statistical methods for analysing primary and secondary outcomes. Reference to where other details of the statistical analysis plan can be found, if not in the protocol	25-30
	20b	Methods for any additional analyses (eg, subgroup and adjusted analyses)	27-28
	20c	Definition of analysis population relating to protocol non-adherence (eg, as randomised analysis), and any statistical methods to handle missing data (eg, multiple imputation)	26
Methods: Monitoring			
Data monitoring	21a	Composition of data monitoring committee (DMC); summary of its role and reporting structure; statement of whether it is independent from the sponsor and competing interests; and reference to where further details about its charter can be found, if not in the protocol. Alternatively, an explanation of why a DMC is not needed	31
	21b	Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial	n.a.
Harms	22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct	n.a.

Auditing	23	Frequency and procedures for auditing trial conduct, if any, and whether the process will be independent from investigators and the sponsor	n.a.
Ethics and dissemination			
Research ethics approval	24	Plans for seeking research ethics committee/institutional review board (REC/IRB) approval	11 and 34
Protocol amendments	25	Plans for communicating important protocol modifications (eg, changes to eligibility criteria, outcomes, analyses) to relevant parties (eg, investigators, REC/IRBs, trial participants, trial registries, journals, regulators)	n.a.
Consent or assent	26a	Who will obtain informed consent or assent from potential trial participants or authorised surrogates, and how (see Item 32)	23
	26b	Additional consent provisions for collection and use of participant data and biological specimens in ancillary studies, if applicable	n.a.
Confidentiality	27	How personal information about potential and enrolled participants will be collected, shared, and maintained in order to protect confidentiality before, during, and after the trial	30-31
Declaration of interests	28	Financial and other competing interests for principal investigators for the overall trial and each study site	33-34
Access to data	29	Statement of who will have access to the final trial dataset, and disclosure of contractual agreements that limit such access for investigators	30-31 and 34
Ancillary and post-trial care	30	Provisions, if any, for ancillary and post-trial care, and for compensation to those who suffer harm from trial participation	23-25
Dissemination policy	31a	Plans for investigators and sponsor to communicate trial results to participants, healthcare professionals, the public, and other relevant groups (eg, via publication, reporting in results databases, or other data sharing arrangements), including any publication restrictions	34
	31b	Authorship eligibility guidelines and any intended use of professional writers	n.a.
	31c	Plans, if any, for granting public access to the full protocol, participant-level dataset, and statistical code	n.a.
Appendices			
Informed consent materials	32	Model consent form and other related documentation given to participants and authorised surrogates	n.a.
Biological specimens	33	Plans for collection, laboratory evaluation, and storage of biological specimens for genetic or molecular analysis in the current trial and for future use in ancillary studies, if applicable	n.a.

*It is strongly recommended that this checklist be read in conjunction with the SPIRIT 2013 Explanation & Elaboration for important clarification on the items. Amendments to the protocol should be tracked and dated. The SPIRIT checklist is copyrighted by the SPIRIT Group under the Creative Commons "Attribution-NonCommercial-NoDerivs 3.0 Unported" license.

A.2 Integrity checklist psychologist/expert-patient sessions

The current session is an EMAIL / CHAT

A. Structure

The following elements should ALL be present in an email or a chat session:

E-mail	**Chat**
1.Extraction of the question 2.Formulation of an answer 3.Ending	1.Warm welcome 2.Establishing the topic 3.Establishing what will be discussed in the current chat 4.Discussing the topic / conveying support or advice 5.Ending
All present? YES / NO	All present? YES / NO

B. Content/Interventions

The aim of the intervention is to make people aware of their eating problems and to provide ways/suggestions to enlarge this insight, counteract eating related problems and/or to stimulate seeking help.

NOTE. If a method/delivery/intervention falls under more categories it only counts as one (no double counts).

The session took at least 20 minutes	YES / NO
The situation was assessed / summarized for participants	YES / NO
The topic of conversation was established	YES / NO
Count the interventions present in the form of:	AMOUNT
1. Giving support / empathy	1
2. Reflecting feelings	2
3. Motivating	3
4. Expressing concern	4
5. Asking for more clarity	5
Count the interventions present in the form of:	AMOUNT
1. Providing Psychoeducation	1
2. Providing advice	2
3. Concretizing aims or goals	3
4. Stimulating thinking / reflection	4
5. Confronting	5
6. Challenging cognitions / beliefs	6
7. Suggesting to seek help / treatment	7
8. Explain procedures	8

9. Other...	9
At the end of the session 1. the participant knows what to do in the short term (coming week) 2. concrete advice or directions are provided by the supporter 3. suggestion(s) about dealing with obstacles or difficulties is/are provided	 YES / NO YES / NO YES / NO
NUMBER OF 'YES' (range 0-6) NUMBER OF INTERVENTIONS (sum of interventions)	

C. Method of Delivery

Contents the way in which interventions are delivered to the participant.

NOTE. If a method/delivery/intervention falls under more categories it only counts as one (no double counts).

A. Interventions were present in the form of: 1. Sharing (common) knowledge or scientific findings to introduce or complement advice or psychoeducation 2. Sharing (common) knowledge or scientific findings to show support or to reduce stigma/feelings of shame ("many people with eating problems...") 3. Presenting solutions to problems by mentioning (directly) that is has been found to work in research or by other people ("... works for many people with eating problems") 4. Sharing one's own experience in a way that a participant feels recognized / to break stigmatization / to give hope to participants 5. Sharing one's own experience to offer advice 6. Sharing one's own experience to stimulate seeking help or treatment 7. Sharing one's own experience to offer psychoeducation	 YES / NO YES / NO YES / NO YES / NO YES / NO YES / NO
B1 At the end of the session it is evident that the supporter is a person with knowledge about the problems the participant is currently struggling with and maintains a psychologist approach B2 At the end of the session it is evident that the supporter has had experience with (a form of) the problems the participant is currently struggling with	YES / NO YES / NO
C1 The supporter never talks about his/her own life (e.g. a situation, feeling/emotion, thought(process), difficulty, success) during the session C2 The supporter has revealed something (e.g. a situation, feeling/emotion, thought(process), difficulty, success) about his/her life as a tool to offer support to the participant during the session (can overlap with the first check of this table)	YES / NO YES / NO
D The supporter never uses medical terminology or medical abbreviations (or if it used: explains what the term/abbreviation means or verifies that the participant knows what the term/abbreviation means)	YES / NO

E1 I (the rater) believe the supporter is a psychologist	YES / NO
E2 I (the rater) believe the supporter is an expert patient	YES / NO

Integrity - Final Score			
A	**Structure**		YES / NO
B	**Content**	Number of interventions used	...
		Number of YES (range 0-6)	...
C	**Method**	Number of YES in A1-3, B1, C1, E1 (range 0-6)	...
		Number of YES in A4-7, B2, C2, D1, E2 (range 0-8)	...

Chapter B

Supplemental Material for Chapter 3

B.1 Overview of statistical models

Condition contrasts (CC)

Condition	CC1	CC2	CC3
Waiting list control	-1	0	0
Featback	1/3	-1/2	-1
Expert-patient support	1/3	-1/2	1
Featback + Expert-patient support (combination)	1/3	1	0

Time contrasts (TC)

Measurement	TC1	TC2	TC3	TC4	TC5
T0, baseline	-1	0	0	0	0
T1, post intervention	1	-1	-1	-1	-1
T2, 3-month FU	0	1	0	0	0
T3, 6-month FU	0	0	1	0	0
T4, 9-month FU	0	0	0	1	0
T5, 12-month FU	0	0	0	0	1

The fifteen possible condition and time contrast combinations were tested separately for the Eating Disorder Examination Questionnaire, General Self-Efficacy Scale, 4-item Patient Health Questionnaire and the 12-item Social Support List (see Table B2.1). Multiple testing was accounted for using a Bonferroni adjustment.

B.2 Results of all tested models

Abbreviations
CC=Condition contrast, CI = confidence interval, TC=Time contrast
CC1 = Three active interventions (Featback only, expert-patient support only and Featback plus expert-patient support) versus waiting list control condition
CC2 = Featback plus expert-patient support condition versus Featback only and expert-patient support only
CC3 = Featback only versus expert-patient support only
TC1 = baseline versus post intervention
TC2 = post intervention versus 3-month follow-up
TC3 = post intervention versus 6-month follow-up
TC4 = post intervention versus 9-month follow-up
TC5 = post intervention versus 12-month follow-up

B2.1 Tested statistical models for the Eating Disorder Examination Questionnaire

CC	TC	Time effects		Time-condition interaction effects		
		β (95% CI)	t (p)	β (95% CI)	t (p)	Cohen's d
1	1	-0.18 (-0.22; -0.14)	-8.12 ($< .001$)	-0.15 (-0.22; -0.07)	-3.66 ($< .001$)	0.38
1	2	-0.12 (-0.18; -0.07)	-4.34 ($< .001$)	0.05 (-0.04; 0.15)	1.05 (.30)	0.11
1	3	-0.16 (-0.22; -0.09)	-4.48 ($< .001$)	0.05 (-0.07; 0.17)	0.88 (.38)	0.11
1	4	-0.23 (-0.31; -0.15)	-5.73 ($< .001$)	0.11 (-0.04; 0.26)	1.49 (.14)	0.20
1	5	-0.27 (-0.35; -0.19)	-6.67 ($< .001$)	0.16 (0.02; 0.29)	2.26 (.02)	0.25
2	1	-0.23 (-0.28; -0.18)	-8.81 ($< .001$)	-0.04 (-0.12; 0.03)	-1.17 (.24)	0.12
2	2	-0.11 (-0.17; -0.04)	-3.35 ($< .001$)	0.01 (-0.08; 0.10)	0.24 (.81)	0.03
2	3	-0.14 (-0.21; -0.06)	-3.57 ($< .001$)	0.02 (-0.09; 0.13)	0.34 (.74)	0.04
2	4	-0.19 (-0.28; -0.11)	-4.28 ($< .001$)	-0.03 (-0.15; 0.10)	-0.41 (.68)	-0.05
2	5	-0.22 (-0.31; -0.13)	-4.87 ($< .001$)	-0.04 (-0.18; 0.09)	-0.64 (.52)	-0.09
3	1	-0.21 (-0.26; -0.15)	-6.73 ($< .001$)	0.01 (-0.06; 0.07)	0.17 (.87)	0.02
3	2	-0.11 (-0.19; -0.03)	-2.78 (.01)	0.06 (-0.02; 0.14)	1.52 (.13)	0.21
3	3	-0.15 (-0.24; -0.06)	-3.13 ($< .001$)	0.02 (-0.07; 0.12)	0.52 (.60)	0.07
3	4	-0.18 (-0.29; -0.07)	-3.33 ($< .001$)	0.06 (-0.05; 0.17)	1.05 (.30)	0.16
3	5	-0.20 (-0.31; -0.09)	-3.66 ($< .001$)	0.01 (-0.09; 0.12)	0.25 (.80)	0.04

B2.2 Tested statistical models for the 4-item Patient Health Questionnaire

CC	TC	Time effects		Time-condition interaction effects		
		β (95% CI)	t (p)	β (95% CI)	t (p)	Cohen's d
1	1	-0.41 (-0.58; -0.23)	-4.61 ($< .001$)	-0.22 (-0.52; 0.08)	-1.43 (.15)	0.12
1	2	-0.08 (-0.25; 0.08)	-1.00 (.32)	0.16 (-0.12; 0.44)	1.11 (.27)	0.11
1	3	-0.07 (-0.26; 0.12)	-0.74 (.46)	-0.07 (-0.39; 0.26)	-0.42 (.68)	0.05
1	4	-0.17 (-0.39; 0.05)	-1.51 (.13)	0.14 (-0.27; 0.54)	0.67 (.51)	0.08
1	5	-0.42 (-0.63; -0.20)	-3.85 ($< .001$)	0.24 (-0.13; 0.60)	1.28 (.20)	0.13
2	1	-0.48 (-0.68; -0.27)	-4.56 ($< .001$)	-0.14 (-0.43; 0.15)	-0.94 (.35)	0.10
2	2	-0.03 (-0.21; 0.16)	-0.30 (.76)	-0.16 (-0.42; 0.11)	-1.14 (.25)	-0.13
2	3	-0.10 (-0.31; 0.12)	-0.87 (.38)	0.22 (-0.07; 0.51)	1.47 (.14)	0.17
2	4	-0.12 (-0.37; 0.13)	-0.95 (.34)	0.08 (-0.28; 0.44)	0.43 (.67)	0.05
2	5	-0.34 (-0.59; -0.09)	-2.67 (.01)	0.10 (-0.24; 0.44)	0.59 (.55)	0.07
3	1	-0.41 (-0.66; -0.15)	-3.14 ($< .001$)	0.13 (-0.13; 0.38)	0.96 (.34)	0.11
3	2	0.05 (-0.17; 0.27)	0.43 (.66)	-0.01 (-0.23; 0.21)	-0.06 (.95)	0.01
3	3	-0.20 (-0.46; 0.05)	-1.6 (.11)	0.10 (-0.15; 0.34)	0.79 (.43)	0.10
3	4	-0.16 (-0.47; 0.15)	-1.03 (.30)	-0.06 (-0.38; 0.26)	-0.36 (.72)	0.05
3	5	-0.39 (-0.69; -0.09)	-2.52 (.01)	-0.20 (-0.5; 0.1)	-1.33 (.19)	0.18

B2.3 Tested statistical models for the General Self-Efficacy Scale

CC	TC	Time effects		Time-condition interaction effects		
		β (95% CI)	t (p)	β (95% CI)	t (p)	Cohen's d
1	1	0.09 (-0.18; 0.35)	0.65 (.52)	0.09 (-0.40; 0.59)	0.37 (.71)	0.04
1	2	0.16 (-0.09; 0.42)	1.23 (.22)	-0.08 (-0.55; 0.40)	-0.32 (.75)	0.04
1	3	0.13 (-0.12; 0.38)	1.02 (.31)	-0.08 (-0.55; 0.39)	-0.34 (.73)	0.04
1	4	0.21 (-0.16; 0.58)	1.10 (.27)	-0.29 (-0.99; 0.42)	-0.80 (.43)	0.12
1	5	0.38 (0.07; 0.69)	2.40 (.02)	-0.15 (-0.73; 0.43)	-0.50 (.62)	0.06
2	1	0.12 (-0.19; 0.43)	0.75 (.45)	-0.14 (-0.59; 0.31)	-0.61 (.54)	0.07
2	2	0.14 (-0.16; 0.43)	0.90 (.37)	-0.15 (-0.56; 0.27)	-0.70 (.48)	0.09
2	3	0.11 (-0.18; 0.40)	0.72 (.47)	-0.20 (-0.62; 0.22)	-0.94 (.35)	0.12
2	4	0.11 (-0.31; 0.54)	0.53 (.59)	-0.32 (-0.90; 0.26)	-1.09 (.28)	0.15
2	5	0.33 (-0.03; 0.70)	1.78 (.08)	-0.32 (-0.85; 0.21)	-1.20 (.23)	0.16
3	1	0.19 (-0.20; 0.58)	0.96 (.34)	-0.14 (-0.52; 0.25)	-0.71 (.48)	0.09
3	2	0.21 (-0.14; 0.57)	1.16 (.25)	-0.07 (-0.44; 0.30)	-0.35 (.72)	0.05
3	3	0.21 (-0.16; 0.57)	1.11 (.27)	0.03 (-0.34; 0.40)	0.17 (.87)	0.02
3	4	0.28 (-0.21; 0.76)	1.12 (.26)	0.08 (-0.40; 0.56)	0.34 (.74)	0.05
3	5	0.50 (0.05; 0.94)	2.20 (.03)	0.18 (-0.25; 0.61)	0.82 (.41)	0.12

B2.4 Tested statistical models for the 12-item Social Support List

CC	TC	Time effects		Time-condition interaction effects		
		β (95% CI)	t (p)	β (95% CI)	t (p)	Cohen's d
1	1	-0.04 (-0.35; 0.26)	-0.28 (.78)	-0.19 (-0.72; 0.35)	-0.69 (.49)	0.07
1	2	-0.16 (-0.49; 0.17)	-0.94 (.35)	0.02 (-0.58; 0.62)	0.07 (.94)	0.01
1	3	-0.17 (-0.54; 0.20)	-0.90 (.37)	0.37 (-0.27; 1.02)	1.14 (.26)	0.13
1	4	-0.10 (-0.57; 0.38)	-0.39 (.70)	0.18 (-0.61; 0.97)	0.44 (.66)	0.06
1	5	0.35 (-0.10; 0.80)	1.52 (.13)	0.20 (-0.58; 0.97)	0.50 (.62)	0.06
2	1	-0.11 (-0.46; 0.24)	-0.60 (.55)	0.17 (-0.33; 0.66)	0.66 (.51)	0.07
2	2	-0.15 (-0.52; 0.23)	-0.77 (.44)	-0.58 (-1.14; -0.02)	-2.05 (.04)	0.29
2	3	-0.04 (-0.46; 0.38)	-0.20 (.84)	-0.27 (-0.93; 0.39)	-0.80 (.43)	0.12
2	4	-0.03 (-0.55; 0.49)	-0.12 (.91)	-0.46 (-1.22; 0.30)	-1.19 (.23)	0.17
2	5	0.42 (-0.09; 0.93)	1.62 (.11)	-0.52 (-1.27; 0.23)	-1.36 (.17)	0.19
3	1	-0.19 (-0.58; 0.20)	-0.95 (.34)	-0.14 (-0.54; 0.26)	-0.69 (.49)	0.09
3	2	0.14 (-0.30; 0.58)	0.63 (.53)	-0.29 (-0.71; 0.14)	-1.31 (.19)	0.18
3	3	0.09 (-0.42; 0.60)	0.35 (.73)	-0.24 (-0.76; 0.28)	-0.91 (.36)	0.13
3	4	0.20 (-0.42; 0.81)	0.63 (.53)	-0.05 (-0.67; 0.57)	-0.16 (.87)	0.03
3	5	0.68 (0.09; 1.27)	2.26 (.02)	-0.23 (-0.79; 0.34)	-0.78 (.44)	0.11

B.3 Reliable Change Index

To calculate the reliable change index (RCI) the method described by Jacobson & Truax (1991) was used.

The standard deviation (SD) of the Eating Disorder Examination Questionnaire (EDE-Q, primary outcome measure) total scores at baseline was 1.04. Cronbach's α between the 22 items that constitute the EDE-Q total score was used as the reliability measure. Cronbach's α was .90. The RCI was calculated with the following formula.

$$RCI = \sqrt{2 * SE^2} * 1.96$$

Where

$$SE = SD * \sqrt{1 - \alpha}$$

The RCI was 0.89. Derived from the RCI, the number of participants (averaged across 100 imputed datasets) showing reliable deterioration, no change and reliable improvement was 14, 261 and 80 respectively. No significant difference in these frequencies between conditions was found, $\chi^2(6) = 11.14$, $p = .08$. Frequencies of reliable deterioration, no change and reliable improvement per condition can be found in Table B3.1. Mean change in EDE-Q scores from baseline to post intervention was 0.38 (pooled $SD = 0.76$), with higher scores meaning more improvement. The mean (SD) in the Featback, combination, expert-patient support and waitlist condition was 0.42 (0.72), 0.54 (0.71), 0.40 (0.85) and 0.06 (0.70) respectively.

B3.1 Frequencies of reliable deterioration, no change and reliable improvement per condition ($N = 355$)

	Featback ($N = 88$)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)
Reliable deterioration	3	3	1	7	14
No change	65	61	65	71	261
Reliable improvement	20	26	21	12	80

Note. Table rows may not add up to the frequencies shown in the 'total sample' column, because of rounding to whole participants across 100 imputed datasets.

Reference

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. https://doi.org/10.1037//0022-006x.59.1.12

B.4 Intervention check results

To verify that support sessions carried out by expert patients were different from those by psychologists, an intervention check was created and rated by two master level psychology students (see study protocol for the intervention check). The two raters evaluated 15 chat and 15 email sessions of psychologist and 15 chat and 15 email sessions of expert patients. Hypotheses in the study protocol concerning the intervention check are addressed here in six separate questions.

Question 1: Could raters distinguish interventions between psychologists and expert patients?
Across the two raters, for 94% of the sessions the supporter was correctly identified as psychologist or expert patient. Agreement between raters was 95%. Additionally, it was hypothesized that methods to convey interventions would differ between psychologists and expert-patients. Indeed, expert-patient deliveries (i.e., interventions that include own experiences) were more frequent than psychologist deliveries (i.e., interventions that include common or scientific knowledge) in sessions executed by expert patients, and less frequent than psychologist deliveries in sessions executed by psychologists (this was true for 95% of the sessions). Similarly, typical psychologist deliveries were more frequent in sessions by psychologists and less frequent in expert-patient sessions compared to expert-patient deliveries (this was true for 95% of the sessions). In other words, expert patients were more likely to convey interventions using their own experiences, while psychologists were more likely to convey interventions using research or (common) knowledge.

Question 2: Is the structure of chats and emails similar between psychologists and expert patients?
The structure of emails and chats as described in the training protocol for psychologists and expert patients (see study protocol) was followed in all sessions (100% agreement between the two raters). Additionally, structure scores (scale 0 – 6) were calculated based on (1) sessions took 20 minutes, (2) situation was assessed or summarized, (3) topic of conversation was established, (4) participant knew what to do in the short term after the session, (5) concrete advice or directions were provided by the supporter, and (6) suggestions to deal with anticipated obstacles to reach the goals were provided. The mean structure score across sessions was 4.8 ($SD = 1.0$) with no difference in structure scores between expert patients and psychologists (pooled mean difference $= 0.1$, pooled $SE = 0.4$, pooled $t(3) = 0.32$, $p = .38$).

Question 3: Do psychologists use a broader pallet of (more distinct) interventions?
On average, supporters used 5.5 (pooled $SD = 1.2$) types of interventions (see integrity checklist from study protocol for the full list of interventions). Furthermore, psychologist used significantly more interventions types (a broader pallet of interventions) than expert patients, pooled mean difference $= 1.1$, pooled $SE = 0.4$, $t(56) = 2.55$, $p < .01$.

Question 4: Do expert patients mention their own experiences in sessions?
Expert patients shared at least some of their own experiences in 93% of the sessions (agree-

ment between raters was 100%).

Question 5: Do psychologists never mention their own experiences in sessions?
Psychologists never shared their own experiences in 100% of the sessions (agreement between raters was 100%).

Question 6: Do psychologists use more medical terms than expert patients?
Medical terms were not used in any of the sessions (100% agreement between raters), regardless of whether the supporter was an expert patient or psychologist.

B.5 Results on intervention usage, satisfaction, help-seeking intentions and behaviors and other e-health use

Assessments will be indicated by the abbreviations T1, T2, T3, T4, T5 for post intervention, 3, 6, 9 and 12-month follow up, respectively. These results are exploratory and based on completed data only. Specifically, intervention usage is based on all participants in the three active intervention conditions; 88, 90 and 87 participants for the Featback, combination and expert-patient support condition respectively. Satisfaction results are based on 212 participants measured at T1. For help-seeking and e-health use results are based on 275 (T1), 249 (T2), 242 (T3), 232 (T4) and 242 (T5) participants.

Intervention usage

In the two conditions where participants could make use of the automated messages system, the average amount of Featback sessions used was 6.03 ($SD = 3.02$, range 0-8). The average amount of used support sessions in the two expert-patient support conditions was 3.97 ($SD = 3.02$, range 0-9). In total, 1074 Featback sessions and 702 support sessions were used. Participants planned 806 sessions, meaning that 104 support sessions did not happen because the participant did not show. Of the 702 used support sessions 368 (52.4%) were via email and 334 (47.6%) via chat, with no differences between the two conditions with expert-patient support in preference for email, $t(175) = 0.47$, $p = .64$, or chat, $t(175) = 1.59$, $p = .12$. Participants in the combined condition with access to both Featback and expert-patient support used significantly less Featback sessions (mean difference = -0.84, $SE = 0.37$; $t(176) = 2.28$, $p = .024$), but not less support sessions (mean difference = -0.83, $SE = 0.45$; $t(175) = 1.86$, $p = .066$) in comparison to the Featback only and expert-patient support only conditions respectively.

Looking at the total amount of sessions used, significant differences were apparent between the intervention conditions ($F(2, 262) = 37.67$, $p < .001$). Participants used the least amount of sessions in the expert-patient support only condition (mean = 4.39, $SD = 3.07$), more in the Featback only condition (mean = 6.45, $SD = 2.09$) and most sessions were used in the combined condition with access to both Featback and expert-patient support (mean = 9.18, $SD = 5.16$), with all comparisons $p < .001$. When looking at intervention usage as percentage of the possible amount of sessions (i.e., 8 in the Featback only and expert-patient support only conditions and 16 in the combined condition) a significant difference was found ($F(2, 262) = 16.80$, $p < .001$) with the Featback only condition having a higher intervention usage (80.7%) compared to the expert-patient support only condition (54.9%), $t(152) = 5.20$, $p < .001$, and the combined condition (57.4%), $t(170) = 5.31$, $p < .001$. There was no apparent difference in this measure of intervention usage between the two conditions with expert-patient support, $t(167.90) = 0.46$, $p = .64$.

Satisfaction

Participants indicated to be satisfied (on a scale from 1 to 10) with the intervention, with a significant difference between the conditions, $F(2, 192) = 15.98$, $p < .001$. The Featback only condition received a lower rating (mean = 5.84, $SD = 1.79$) than the combined

condition (mean = 7.06, $SD = 1.69$), $t(131) = 4.03$, $p < .001$, and the expert-patient only condition (mean = 7.43, $SD = 1.63$), $t(130) = 5.34$, $p < .001$, with no difference between the expert-patient support only and combined conditions, $t(123) = 1.26$, $p = .21$. The total amount of completed Featback and/or expert-patient sessions correlated positively with satisfaction ratings ($r = .18$; $\beta = 0.09$, $t(193) = 2.61$, $p = .010$, indicating that completing more sessions was associated with higher satisfaction. Further exploration revealed no differences in intervention satisfaction between participants who indicated never to have had eating disorder related treatment ($N = 78$, mean = 6.7, $SD = 1.7$) and those who did ($N = 117$, mean = 6.8, $SD = 1.9$), $t(193) = 0.16$, $p = .87$. Details per condition are presented below.

Featback only condition ($N = 72$)
Overall grade (scale of 1 lowest – 10 highest): 5.8 ($SD = 1.8$).
Participants were neutral about the quality ($M = 3.0$, $SD = 0.8$; scale 1 excellent – 4 poor) and received support ($M = 4.1$, $SD = 1.5$; scale 1 very dissatisfied – 7 very satisfied) of Featback. The majority of participants ($n = 46$, 63.9%) in this condition did not learn new things from the program, but it did help to make participants ($n = 42$, 58.3%) more aware of their problems. Sixty (83.3%) of the participants thought the feedback of the messages was at least moderately applicable. All participants thought the idea of individual monitoring to be good. The most useful features in this condition were rated to be the weekly feedback on well-being, the tips and advice in the Featback messages and the feeling of working towards recovery. Free text on negative aspects of Featback concerned (1) the program being too shallow or focusing too little on underlying mechanisms, (2) the program not being intensive/long enough, (3) missing personal contact/someone to talk to, (4) messages not being applicable or useful. Positive aspects of Featback concerned (1) making one aware of one's problems, (2) motivating messages containing diverse and useful advice, (3) being low-threshold and a good first step.

Featback and expert-patient support ($N = 72$)
Overall grade (scale of 1 lowest – 10 highest): 7.1 ($SD = 1.7$).
Participants in this condition were satisfied with the quality ($M = 2.6$, $SD = 0.9$; scale 1 excellent – 4 poor) and received support (M = 5.1, SD = 1.4; scale 1 very dissatisfied – 7 very satisfied). The majority of participants (n = 48, 66.7%) learned new things from the intervention and it helped to make participants (n = 53, 73.6%) more aware of their problems. Sixty-four (88.9%) of the participants thought the feedback of the automated messages was at least moderately applicable. All participants thought the idea of individual monitoring to be good and almost all participants (63, 87.5%) thought the idea of extra individual support of an expert patient to be good. The most useful feature in this condition was rated to be the expert-patient support. Other useful features were the weekly feedback on wellbeing, the tips and advice in the Featback messages and the feeling of working towards recovery. Negative and positive aspects of the Featback messages were similar to those described before. Negative aspects of the expert-patient support concerned (1) 20-minute chats were too short, (2) no match regarding the eating disorder between participant and supporter, (3) technical problems with making appointments or chats. Positive aspects of the expert-patient support concerned (1) the warm and empathic approach of expert pa-

tients, (2) having someone to talk to/being recognized/not feeling alone or crazy, (3) honest and open feedback that was applicable, (4) gaining insight.

Expert-patient support ($N = 68$)
Overall grade (scale of 1 lowest – 10 highest): 7.4 ($SD = 1.6$).
Positive and negative aspects of expert-patient support were similar to those mentioned before. The main negative aspect was that many participants argued that 20-minute chat sessions were very brief. The main positive aspect was that many participants felt recognized and easily understood, which was the basis for fruitful contact with useful advice.

Help seeking intentions and behaviors

Help seeking intentions and behaviors and the influence of the active interventions were inquired. Participants reported (on a 7-point scale) to believe to be in need of help at T1 ($M = 5.6$, $SD = 1.5$) and across T1-T5 ($M = 5.0$, pooled $SD = 1.8$) and even more in need of professional help at T1 ($M = 5.9$, $SD = 1.4$) and across T1-T5 ($M = 5.7$, pooled $SD = 1.5$). At T1-T5 150 (54.5%), 156 (62.7%) 140 (57.9%) 125 (53.9%) 130 (36.6%) participants indicated to have pursued professional help because of disordered eating or body dissatisfaction. The majority went to treatment facilities or a psychologist. Of the participants who pursued professional help 22.0% (T1), 12.8% (T2), 16.3% (T3), 15.9% (T4) and 16.9% (T5) indicated that the intervention stimulated them to take this step.

Some participants (ranging from 10/108 (9.3%) at T4 and 28/130 (21.5%) at T1) did not seek out professional help, but did have intentions to do so. Of these participants 57.7% (T1), 81.8% (T2), 38.5% (T3), 55.6% (T4) and 42.9% (T5) indicated that the intervention stimulated them to form the intention to pursue professional help.

However, a majority of participants (ranging from 102/130 (78.5%) at T1 to 98/108 (90.7%) at T4) who did not pursue professional help also did not have intentions to seek help. Most named reasons for not pursuing help were (1) not considering the problems to be serious enough, (2) not wanting others to find out about the eating problem, (3) feelings of shame, (4) fear of stigmatization, (5) worries about costs, (6) not knowing where to go and (7) believing that health professionals will only make it worse or will not understand or take it seriously.

eHealth use

The majority of participants, specifically 193 (70.2%), 191 (76.7%), 161 (66.7%), 150 (64.7%) and 148 (61.2%) at T1, T2, T3, T4 and T5, indicated to have used websites beside Featback in relation to their eating problems. The number of participants using a forum declined from 134 (48.7%) at T1 to 70 (28.9%) at T5, with the majority of participants indicating not to have used a forum, 141 (51.3%) at T1 and 172 (71.1%) at T5. Of the participants who indicated to have used a forum around one third actively participated by posting content themselves; on average 31.1% (range 26.3%-35.7%) at each assessment. Lastly, participants were asked whether they have made use of another email or chat service for their eating problem. On average 15.2% (range 12.0%-18.1%) of participants at each assessment had made use of another email or chat service, mostly provided by the Proud2Bme community.

Chapter C

Supplemental Material for Chapter 4

C.1 Prisma 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review	- Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	- Abstract
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	- Introduction
Rationale	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	- Pooling cost-effectiveness data: last paragraph
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	- Eligibity criteria
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	- Search strategy and selection criteria
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	- Search strategy and selection criteria - Multimedia Appendix
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	- Search strategy and selection criteria: last paragraph
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	- Search strategy and selection criteria: last paragraph

Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	- Data preparation - Multimedia Appendix
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	- Data preparation - Multimedia Appendix
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	- Quality assessment
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	- Statistical analyses - Table 1
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	- Table 2 and 3 - Statistical analyses: moderators
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	- Data preparation - Table 1
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	- Table 2 and 3 - Statistical analyses: heterogeneity and publication bias
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	- Statistical analyses - Statistical analyses: pooling incremental net benefits - Statistical analyses: heterogeneity and publication bias
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	- Statistical analyses: moderators
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	- Statistical analyses: sensitivity analyses
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	- Quality assessment - Statistical analyses: sensitivity analyses
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	- Statistical analyses: sensitivity analyses
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	- Figure 1 - Table 2 and 3 - Characteristics of included studies
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	- Multimedia Appendix

Study characteristics	17	Cite each included study and present its characteristics.	- Table 2 and 3
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	- Table 4 - Table 5
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	- Table 2 and 3
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	- Risk of bias and quality of economic evaluation - Sensitivity analyses
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	- Quality of life - Costs - Cost-effectiveness - Moderator analyses - Sensitivity analyses
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	- Cost-effectiveness - Publication bias
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	- Sensitivity analyses
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	- Sensitivity analyses
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed. 95% Cis are presented.	- Quality of life - Costs - Cost-effectiveness - Moderator analyses - Sensitivity analyses
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	- Discussion: first paragraph
	23b	Discuss any limitations of the evidence included in the review.	- QALYs and mental health interventions - Meta-analyses on cost-effectiveness data - Limitations
	23c	Discuss any limitations of the review processes used.	- Limitations
	23d	Discuss implications of the results for practice, policy, and future research.	- Discussion: first and second paragraph - Future directions
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	- Abstract - Methods
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	- Abstract - Methods
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	- Deviations from the protocol
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	- Acknowledgements
Competing interests	26	Declare any competing interests of review authors.	- Conflicts of interest

Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	All available upon reasonable request - Multimedia Appendix

C.2 Full Pubmed search string

(mental disorders[Mesh] OR mental illness[tiab] OR mental disorder[tiab] OR psychological disorder[tiab] OR psychological illness[tiab] OR psychological disease[tiab] OR psychiatric disorder[tiab] OR psychiatric illness[tiab] OR psychiatric disease[tiab] OR mood[tiab] OR anxiety[tiab] OR anxiety disorders[tiab] OR agoraphobia[tiab] OR separation anxiety[tiab] OR neurocirculatory Asthenia[tiab] OR neurotic Disorders[tiab] OR obsessive-compulsive[tiab] OR hoarding disorder[tiab] OR panic disorder[tiab] OR phobic disorders[tiab] OR social phobia[tiab] OR bipolar and related disorders[tiab] OR bipolar[tiab] OR disruptive disorders[tiab] OR impulse control[tiab] OR conduct disorders[tiab] OR firesetting behavior[tiab] OR gambling[tiab] OR trichotillomania[tiab] OR dissociative disorders[tiab] OR dissociative identity disorder[tiab] OR elimination disorders[tiab] OR encopresis[tiab] OR enuresis[tiab] OR diurnal enuresis[tiab] OR nocturnal enuresis[tiab] OR feeding and eating disorders[tiab] OR eating disorders[tiab] OR anorexia nervosa[tiab] OR binge-eating[tiab] OR bulimia nervosa[tiab] OR feeding and eating disorders of childhood[tiab] OR female athlete triad syndrome[tiab] OR food addiction[tiab] OR night eating[tiab] OR pica[tiab] OR mood disorders[tiab] OR cyclothymic disorder[tiab] OR depressive disorder[tiab] OR depression[tiab] OR postpartum depression[tiab] OR major depressive disorder[tiab] OR treatment-resistant depressive disorder[tiab] OR dysthymic disorder[tiab] OR premenstrual dysphoric disorder[tiab] OR seasonal affective disorder[tiab] OR motor disorders[tiab] OR neurocognitive disorders[tiab] OR amnesia[tiab] OR anterograde amnesia[tiab] OR retrograde amnesia[tiab] OR transient global amnesia[tiab] OR cognition disorders[tiab] OR auditory perceptual disorders[tiab] OR cognitive dysfunction OR Huntington disease[tiab] OR consciousness disorders[tiab] OR delirium[tiab] OR emergence delirium[tiab] OR dementia[tiab] OR AIDS dementia complex[tiab] OR Alzheimer[tiab] OR primary progressive aphasia[tiab] OR primary progressive nonfluent aphasia[tiab] OR Creutzfeldt-Jakob[tiab] OR vascular dementia[tiab] OR multi-infarct dementia[tiab] OR diffuse neurofibrillary tangles with calcification[tiab] OR frontotemporal lobar degeneration[tiab] OR frontotemporal dementia[tiab] OR Pick disease of the brain[tiab] OR Kluver-Bucy syndrome[tiab] OR Lewy Body disease[tiab] OR acquired dyslexia[tiab] OR pure alexia[tiab] OR neurodevelopmental disorders[tiab] OR attention deficit and disruptive behavior disorders[tiab] OR ADHD[tiab] OR ADD[tiab] OR attention deficit disorder with hyperactivity[tiab] OR conduct disorder[tiab] OR child behavior disorders[tiab] OR pervasive child development disorders[tiab] OR autism[tiab] OR Asperger syndrome[tiab] OR autistic[tiab] OR communication disorders[tiab] OR language disorders[tiab] OR agraphia[tiab] OR anomia[tiab] OR dyslexia[tiab] OR language development disorders[tiab] OR speech disorders[tiab] OR aphasia[tiab] OR Broca aphasia[tiab] OR conduction aphasia[tiab] OR primary progressive aphasia[tiab] OR primary progressive nonfluent aphasia[tiab] OR Wernicke aphasia[tiab] OR articulation disorders[tiab] OR dysarthria[tiab] OR echolalia[tiab] OR mutism[tiab] OR stuttering[tiab] OR learning disorders[tiab] OR dyscalculia[tiab] OR acquired dyslexia[tiab] OR developmental disabilities[tiab] OR intellectual disability[tiab] OR learning disorders[tiab] OR motor skills disorders[tiab] OR mutism[tiab] OR reactive attachment disorder[tiab] OR childhood schizophrenia[tiab] OR stereotypic movement disorder[tiab] OR tic disorders[tiab] OR Tourette syndrome[tiab] OR neurotic disorders[tiab] OR paraphilic disorders[tiab] OR exhibitionism[tiab] OR fetishism[tiab] OR masochism[tiab] OR pedophilia[tiab] OR sadism[tiab] OR transvestism[tiab] OR voyeurism[tiab] OR personality

disorder[tiab] OR antisocial personality disorder[tiab] OR borderline personality disorder[tiab] OR compulsive personality disorder[tiab] OR dependent personality disorder[tiab] OR histrionic personality disorder[tiab] OR hysteria[tiab] OR paranoid personality disorder[tiab] OR passive-aggressive personality disorder[tiab] OR schizoid personality disorder[tiab] OR schizotypal personality disorder[tiab] OR schizophrenia spectrum and other psychotic disorders[tiab] OR schizophrenia spectrum[tiab] OR psychotic affective disorders[tiab] OR Capgras syndrome[tiab] OR delusional parasitosis[tiab] OR Morgellons disease[tiab] OR paranoid disorders[tiab] OR psychotic disorders[tiab] OR substance-induced psychoses[tiab] OR alcoholic psychoses[tiab] OR schizophrenia[tiab] OR catatonic schizophrenia[tiab] OR disorganized schizophrenia[tiab] OR paranoid schizophrenia[tiab] OR shared paranoid disorder[tiab] OR psychological sexual dysfunctions[tiab] OR dyspareunia[tiab] OR erectile dysfunction[tiab] OR gender dysphoria[tiab] OR premature ejaculation[tiab] OR sexual and gender disorders[tiab] OR vaginismus[tiab] OR sleep wake disorders[tiab] OR dyssomnias[tiab] OR sleep deprivation[tiab] OR circadian rhythm sleep disorders[tiab] OR jet lag syndrome[tiab] OR intrinsic sleep disorders[tiab] OR disorders of excessive somnolence[tiab] OR idiopathic hypersomnolence[tiab] OR Kleine-Levin syndrome[tiab] OR narcolepsy[tiab] OR cataplexy[tiab] OR restless legs syndrome[tiab] OR sleep initiation and maintenance disorders[tiab] OR parasomnias[tiab] OR nocturnal myoclonus syndrome[tiab] OR nocturnal paroxysmal dystonia[tiab] OR REM sleep parasomnias[tiab] OR REM sleep behavior disorder[tiab] OR sleep paralysis[tiab] OR sleep arousal disorders[tiab] OR night terrors[tiab] OR somnambulism[tiab] OR sleep bruxism[tiab] OR sleep-wake transition disorders[tiab] OR somatoform disorders[tiab] OR body dysmorphic[tiab] OR conversion disorder[tiab] OR factitious disorders[tiab] OR Munchausen syndrome[tiab] OR Munchausen syndrome by proxy[tiab] OR hypochondriasis[tiab] OR neurasthenia[tiab] OR substance-related disorders[tiab] OR addiction[tiab] OR alcohol-related disorders[tiab] OR alcohol amnestic disorder[tiab] OR alcoholic Korsakoff syndrome[tiab] OR alcohol withdrawal delirium[tiab] OR alcoholic intoxication[tiab] OR alcoholism [tiab] OR binge drinking[tiab] OR alcoholic psychoses[tiab] OR Wernicke encephalopathy[tiab] OR amphetamine-related disorders[tiab] OR cocaine-related disorders[tiab] OR inhalant abuse[tiab] OR marijuana abuse[tiab] OR marijuana use[tiab] OR neonatal abstinence syndrome[tiab] OR opioid-related disorders[tiab] OR heroin dependence[tiab] OR morphine dependence[tiab] OR opium dependence[tiab] OR phencyclidine abuse[tiab] OR substance-induced psychoses[tiab] OR intravenous substance abuse[tiab] OR oral substance abuse[tiab] OR substance withdrawal syndrome[tiab] OR tobacco use disorder[tiab] OR trauma and stressor related disorders[tiab] OR adjustment disorders[tiab] OR traumatic stress disorders[tiab] OR battered child syndrome[tiab] OR combat disorders[tiab] OR psychological trauma[tiab] OR post-traumatic stress[tiab] OR acute traumatic stress[tiab] OR PTSD[tiab]) **AND** ("telemedicine"[Mesh] OR "telenursing"[Mesh] OR "user-computer interface"[Mesh] OR "multimedia"[Mesh] OR "cell phone"[Mesh] OR "public health informatics"[Mesh] OR "medical informatics"[Mesh] OR "nursing informatics"[Mesh] OR "computers, handheld"[Mesh] OR "mobile applications"[Mesh] OR "internet"[Mesh] OR "patient portals"[Mesh] OR econsult*[tiab] OR e-treat*[tiab] OR e-therap*[tiab] OR e-consult*[tiab] OR ediagnos*[tiab] OR e diagnos*[tiab] OR mobile health*[tiab] OR mhealth*[tiab] OR m health*[tiab] OR telehealth*[tiab] OR tele health[tiab] OR remote consult*[tiab] OR teleconsult*[tiab] OR tele consult*[tiab] OR telenursing[tiab] OR tele nursing[tiab] OR telediagnos*[tiab] OR tele diagnos*[tiab] OR telemedic*[tiab] OR tele medic*[tiab] OR telemon-

itor*[tiab] OR tele monitor*[tiab] OR ehealth*[tiab] OR e-health*[tiab] OR telecare[tiab] OR tele care[tiab] OR digital health[tiab] OR app[tiab] OR apps[tiab] OR smartphone*[tiab] OR phone application*[tiab] OR telephone application*[tiab] OR mobile application*[tiab] OR mobile technolog*[tiab] OR health technolog*[tiab] OR health application*[tiab] OR internet*[tiab] OR world wide web*[tiab] OR webportal*[tiab] OR web portal*[tiab] OR patient portal*[tiab] OR ipad[tiab] OR ipads[tiab] OR sms[tiab] OR mms[tiab] OR text messag*[tiab] OR ussd[tiab] OR pda[tiab] OR laptop*[tiab] OR palmtop*[tiab] OR palm top*[tiab] OR personal digital assistant*[tiab] OR telecounsel*[tiab] OR tele counsel*[tiab] OR remote counsel*[tiab] OR distance consult*[tiab] OR distance counsel*[tiab] OR distant consult*[tiab] OR patient monitoring[tiab] OR interactive voice response*[tiab] OR multimedia[tiab] OR Mhapps[tiab] OR iphone*[tiab] OR android[tiab] OR game*[tiab] OR gaming[tiab] OR gamification[tiab] OR whatsapp*[tiab] OR e-coach*[tiab] OR wearable*[tiab] OR social media[tiab] OR online*[tiab] OR computer*[tiab] OR electronic*[tiab] OR digital*[tiab] OR "online social network"[tiab] OR "online social networks"[tiab] OR facebook[tiab] OR exergam*[tiab] OR serious gam*[tiab] OR personal health record*[tiab] OR personal electronic health record*[tiab] OR health kiosk*[tiab] OR internet-based[tiab] OR internet based[tiab] OR web-based[tiab] OR web based[tiab] OR iCBT[tiab] OR oCBT[tiab] OR teleconferenc*[tiab] OR tele conferenc*[tiab] OR tele-conferenc*[tiab] OR telephone*[tiab] OR e-counsel*[tiab] OR short message service[tiab] OR SMS[tiab] OR cell-phone[tiab] OR cellphone[tiab] OR cellular phone*[tiab] OR blended*[tiab] OR email*[tiab] OR e-mail*[tiab] OR video-guid*[tiab] OR videoguid*[tiab] OR video-mediated[tiab] OR video-based[tiab] OR videobased[tiab] OR video-deliver*[tiab] OR video-treat*[tiab] OR video-therap*[tiab] OR videothera*[tiab] OR video-intervention*[tiab] OR video-counsel*[tiab] OR video-assist*[tiab] OR video-conferenc*[tiab] OR videoconferenc*[tiab] OR video-monit*[tiab] OR videomonit*[tiab] OR video-communicat*[tiab] OR videocommunicat*[tiab] OR video-remind*[tiab] OR video-administered*[tiab] OR video-aided[tiab] OR video-application*[tiab] OR video-consult*[tiab] OR videoconsult*[tiab] OR video-enabled[tiab] OR Twitter[tiab] OR Facebook[tiab] OR Instagram[tiab] OR forum[tiab] OR chat*[tiab] OR virtual reality*[tiab] OR virtual-reality*[tiab] OR avatar*[tiab] OR Conversational agent*[tiab] OR virtual coach[tiab] OR virtual agent*[tiab] OR embodied agent*[tiab] OR avatar*[tiab] OR relational agent*[tiab] OR interactive agent*[tiab] OR virtual character*[tiab] OR virtual human*[tiab] OR virtual assistant*[tiab] OR telepsychiatry[tiab] OR telepsychiatry[tiab] OR tele-guid*[tiab] OR teleguid*[tiab] OR tele-based[tiab] OR tele-deliver*[tiab] OR teledeliver*[tiab] OR tele-treat*[tiab] OR teletreat*[tiab] OR tele-therap*[tiab] OR telethera*[tiab] OR tele-intervention*[tiab] OR tele-assist*[tiab] OR tele-communicat*[tiab] OR telecommunicat*[tiab]) **AND** (cost-benefit analysis[Mesh] OR "cost effectiveness analysis"[tiab] OR "cost effectiveness analyses"[tiab] OR "cost effectiveness"[tiab] OR "cost effective"[tiab] OR "economic evaluation"[tiab] OR "economic evaluations"[tiab] OR "cost benefit"[tiab] OR "cost-benefit analysis"[tiab] OR "cost-benefit analyses"[tiab] OR "cost-benefit data"[tiab] OR "cost utility"[tiab] OR "cost-utility analysis"[tiab] OR "cost-utility analyses"[tiab] OR marginal analyses[tiab] OR marginal analysis[tiab] OR cost minimization[tiab] OR cost-minimization[tiab] OR cost impact[tiab] OR cost-impact[tiab] OR budget impact[tiab] OR budget-impact[tiab])

C.3 Data extraction items

Category	Extracted item
General	•Author •Year of publication •Journal •Country •Randomized controlled trial (yes/no)
Participants	•Recruitment (com=community/open/mass media; clin=clinical recruitment; scr=systematic screening of a predefined population; other) •Sample size (total) •Sample size (per condition) •% female •Mean age (standard deviation) •Targeted mental disorder •Diagnose (1=formal diagnosis; 2=self-report; 3=other, please specify) •Instrument for diagnosis •Inclusion criteria •Exclusion criteria
Interventions	•Intervention frequency and duration (per condition) •Follow-up (i.e., time between baseline and last follow-up assessment) •Assessment time points of quality of life/utility •Assessment time points of costs/health care use •Intervention description (per condition) •Type of guidance (1=fully automated or no guidance; 2=asynchronous or guidance not at the same time such as e-mail/written feedback; 3=synchronous or guidance at the same time such as chat, telephone and face-to-face) •Intensity of guidance (0=less than once a week; 1=once a week; 2=more than once a week; 3=self-guided; 4=other, please specify)
Questionnaires and methods	•Intention to treat analyses (yes/no) •Primary outcome + measurement instrument •Instrument used for quality of life/utility •Instrument used for costs/health care use •Source of (health care) unit costs •Currency + year of indexing •Discounting (0=none; 1=cost and effects at the same percentage, please specify %; 2=costs and effects at different percentages, please specify both percentages) •Perspective used (1=health care; 2=societal; 3=other, please specify) •If societal perspective was used, method for assessing productivity losses (1=friction cost; 2=human capital; 3=other, please specify)

Outcomes	•QALYs (per condition) •SD/SE/variance/confidence interval of QALYs (per condition) •Delta QALY (i.e., QALYs intervention – QALYs control) •SD/SE/variance/confidence interval of delta QALY •Health care costs (per condition) •SD/SE/variance/confidence interval of health care costs (per condition) •Delta health care costs (i.e., health care costs intervention – health care costs control) •SD/SE/variance/confidence interval of delta health care costs •Societal costs (per condition) •SD/SE/variance/confidence interval of societal costs (per condition) •Delta societal costs (i.e., societal costs intervention – societal costs control) •SD/SE/variance/confidence interval of delta societal costs •Incremental cost-effectiveness ratio •Confidence interval of incremental cost-effectiveness ratio •Cost-effectiveness plane for delta costs and delta QALYs provided (0=no; 1=yes)
Risk of bias	•Random sequence generation •Allocation concealment •Blinding of participants and personnel •Blinding of outcome assessors •Incomplete outcome data •Selective reporting •Other sources of bias

Quality of economic evaluation (CHEC)	•Is the study population clearly described? •Are competing alternatives clearly described? •Is a well-defined research question posed in answerable form? •Is the economic study design appropriate to the stated objective? •Is the chosen time horizon appropriate to include relevant costs and consequences? •Is the actual perspective chosen appropriate? •Are all important and relevant costs for each alternative identified? •Are all costs measured appropriately in physical units? •Are costs valued appropriately? •Are all important and relevant outcomes for each alternative identified? •Are all outcomes measured appropriately? •Are outcomes valued appropriately? •Is an incremental analysis of costs and outcomes of alternatives performed? •Are all future costs and outcomes discounted appropriately? •Are all important variables, whose values are uncertain, appropriately subjected to sensitivity analysis? •Do the conclusions follow from the data reported? •Does the study discuss the generalizability of the results to other settings and patient/client groups? •Does the article indicate that there is no potential conflict of interest of study researcher(s) and funder(s)? •Are ethical and distributional issues discussed appropriately?

SD=standard deviation; SE=standard error; QALY=quality-of-life adjusted life year

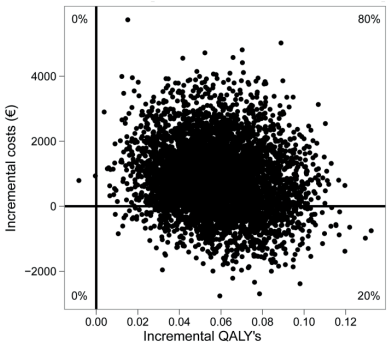
C.4 Formulas used for data preparation and final analyses

Outcome	Nr.	Formula
Data preparation		
	1	$Var_{\Delta QALY} = SD^2_{\Delta QALY}$
	2	$Var_{\Delta QALY} = SE^2_{\Delta QALY}$
	3	$SE_{\Delta QALY} = \frac{UL - Mean_{\Delta QALY}}{1.96}$
	4	$Var_{\Delta QALY} = \frac{SD^2_{intervention}}{N_{intervention}} + \frac{SD^2_{control}}{N_{control}}$
	5	$Var_{\Delta QALY} = SE^2_{intervention} + SE^2_{control}$
	6	$Covariance_{(\Delta QALY, \Delta Costs)} = SD_{\Delta QALY} * SD_{\Delta Costs} * r_{(\Delta QALY, \Delta Costs)}$
	7	$INB_{study} = k * \Delta QALY - \Delta Costs$ Where k is society's willingness to pay for one QALY
Pooling studies		
	8	$INB_{pooled} = \sum(weight_{study} * INB_{study})$
	9	$weight_{study} = \frac{1}{Var(INB_{study}) + \tau^2}$
	10	$Var(INB_{study}) = \frac{1/(k^2 * Var_{\Delta QALY} + Var_{\Delta Costs} - 2 * k * Covariance_{(\Delta QALY, \Delta Costs)}}{\sum(1/Var(INB_{study})}$
	11	$\tau^2 = \frac{Q-(s-1)}{\sum(weight_{study}) * \frac{\sum(weight_{study})}{\sum(weight_{study})^2}}$ Where s is the number of included studies or comparisons, Q is the Cochran statistic and $\tau^2 = 0$ if $Q < s - 1$
Heterogeneity		
	12	$CochranQ = \sum_{s-1}^{s} \frac{1}{Var(INB_{study})} * (INB_{study} - INB_{pooled})^2$ Where $Q = 0$ if $Q < s - 1$
	13	$I^2 = \frac{Q-(s-1)}{Q} * 100\%$

Note. The formula used to calculate delta QALY can also be used to calculate delta costs.
INB=incremental net benefit; Nr=reference number; QALY=quality-of-life-adjusted life year; SD=standard deviation; SE=standard error; UL=upper limit; Var=variance

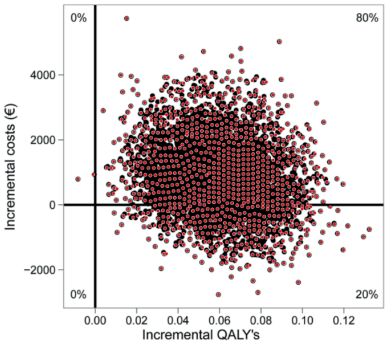
C.5 Example of the estimation of the covariance between delta QALY and delta costs with the use of Webplot Digitizer

[Step 1]
Save the target cost-effectiveness plane with delta QALYs (x-axis) and delta costs (y-axis) as an image.



[Step 2]
Upload the image to Webplot Digitizer (https://automeris.io/WebPlotDigitizer/).
Calibrate x and y values, and select areas to reverse engineer individual data points.

[Step 3]
Run the application and download individual data points into a .csv format.



[Step 4]
Calculate the covariance and, if necessary, the standard deviations of delta QALY and delta costs using the appropriate Excel functions: `=COVARIANCE.S(range delta QALYs; range delta costs)` and `=STDEV.S(range)`.

C.6 Excluded studies with reasons for exclusion

#	Author (year of publication)	Title	Reason for exclusion
1	Adewuya et al. (2019)	The effectiveness and acceptability of mobile telephone adherence support for management of depression in the Mental Health in Primary Care (MeHPriC) project, Lagos, Nigeria: A pilot cluster randomised controlled trial	Wrong outcomes
2	Andersson et al. (2011)	Cost-effectiveness of internet-based cognitive behavior therapy for irritable bowel syndrome: results from a randomized controlled trial	Wrong patient population
3	Andersson et al. (2015)	Cost-effectiveness of an internet-based booster program for patients with obsessive-compulsive disorder: Results from a randomized controlled trial	Wrong outcomes
4	Andersson et al. (2015)	Cost-effectiveness of internet-based cognitive behavior therapy for obsessive-compulsive disorder: Results from a randomized controlled trial	Wrong outcomes
5	Angus et al. (2019)	Cost-effectiveness of strategies to improve delivery of brief interventions for heavy drinking in primary care: results from the ODHIN trial	Wrong study design
6	Axelsson et al. (2018)	Cost-effectiveness and long-term follow-up of three forms of minimal-contact cognitive behaviour therapy for severe health anxiety: Results from a randomised controlled trial	Wrong outcomes
7	Bergstrom et al. (2010)	Internet-versus group-administered cognitive behaviour therapy for panic disorder in a psychiatric setting: a randomised trial	Wrong outcomes
8	Bischof et al. (2010)	Stepped-care intervention for alcohol problems: A cost-effective approach for brief interventions in primary care?	Study protocol
9	Blankers et al. (2012)	Clinical outcomes and economic evaluation of internet-based interventions for harmful alcohol use: a pragmatic randomized trial	Duplicate study
10	Blankers et al. (2012)	Economic evaluation of internet-based interventions for harmful alcohol use alongside a pragmatic randomized controlled trial	Wrong compara-tor
11	Boege et al. (2015)	Cost-effectiveness of intensive home treatment enhanced by inpatient treatment elements in child and adolescent psychiatry in Germany: A randomised trial	Wrong outcomes
12	Bogosian et al. (2021)	Acceptability and Feasibility of a Mindfulness Intervention Delivered via Videoconferencing for People With Parkinson's	Wrong outcomes
13	Bolier et al. (2014)	Cost-effectiveness of online positive psychology: Randomized controlled trial	Wrong outcomes
14	Botha et al. (2018)	Brief Report: A Randomized Control Trial Assessing the Influence of a Telephone-based Intervention on Readmissions for Patients with Severe Mental Illness in a Developing Country	Wrong outcomes
15	Brabyn et al. (2016)	The second Randomised Evaluation of the Effectiveness, cost-effectiveness and Acceptability of Computerised Therapy (REEACT-2) trial: does the provision of telephone support enhance the effectiveness of computer-delivered cognitive behaviour therapy? A randomised controlled trial	Wrong compara-tor
16	Budney et al. (2015)	Computer-assisted behavioral therapy and contingency management for cannabis use disorder	Wrong outcomes

17	Calhoun et al. (2016)	Comparative effectiveness of an Internet-based smoking cessation intervention versus clinic-based specialty care for veterans	Wrong outcomes
18	Celano et al. (2015)	Cost-effectiveness of a collaborative care depression and anxiety treatment program in patients with acute cardiac illness	Wrong intervention
19	Chalder et al. (2012)	A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: the treating depression with physical activity (TREAD) trial	Wrong intervention
20	Chan et al. (2008)	Depression and comorbid PTSD in veterans: Evaluation of collaborative care programs and impact on utilization and costs	Wrong outcomes
21	ChoiYoo et al. (2014)	Cost effectiveness of telecare management for pain and depression in patients with cancer: results from a randomized trial	Wrong outcomes
22	Compen et al. (2017)	Face-to-face versus individual internetbased MBCT versus TAU for distressed cancer patients: The BeMind study	Conference abstract
23	Crow et al. (2009)	The cost effectiveness of cognitive behavioral therapy for bulimia nervosa delivered via telemedicine versus face-to-face	Wrong outcomes
24	Davidson et al. (2013)	Centralized, Stepped, Patient Preference-Based Treatment for Patients With Post-Acute Coronary Syndrome Depression CODIACS Vanguard Randomized Controlled Trial	Wrong outcomes
25	De Boer et al. (2014)	A randomized controlled trial of an Internet-based cognitive-behavioural intervention for non-specific chronic pain: An effectiveness and cost-effectiveness study	Wrong outcomes
26	De Bruin et al. (2016)	Cost-Effectiveness of Group and Internet Cognitive Behavioral Therapy for Insomnia in Adolescents: Results from a Randomized Controlled Trial	Wrong comparator
27	De Graaf et al. (2011)	One-year follow-up results of unsupported online computerized cognitive behavioural therapy for depression in primary care: A randomized trial	Wrong outcomes
28	Dear et al. (2020)	A Cost-effectiveness Analysis of an Internet-delivered Pain Management Program Delivered With Different Levels of Clinician Support: Results From a Randomised Controlled Trial	Wrong patient population
29	Delgadillo et al. (2017)	Improving the efficiency of psychological treatment using outcome feedback technology	Wrong study design
30	Dieng et al. (2013)	A randomised controlled trial of a psycho-educational intervention for melanoma survivors at high risk of developing new primary disease	Study protocol
31	Donohue et al. (2012)	12-Month cost-effectiveness of telephonedelivered collaborative care for treating post-CABG depression	Conference abstract
32	Donohue et al. (2014)	Twelve-month cost-effectiveness of telephone-delivered collaborative care for treating depression following CABG surgery: a randomized controlled trial	Wrong intervention
33	Dorstyn et al. (2012)	Effectiveness of telephone counseling in managing psychological outcomes after spinal cord injury: a preliminary study	Wrong outcomes
34	Downe-Wamboldt et al. (2007)	The effects and expense of augmenting usual cancer clinic care with telephone problem-solving counseling	Wrong outcomes
35	Drost et al. (2016)	A Web-Based Computer-Tailored Alcohol Prevention Program for Adolescents: Cost-Effectiveness and Intersectoral Costs and Benefits	Wrong outcomes

36	Duarte et al. (2014)	Computerised Cognitive Behaviour Therapy for Depression Management: A Cost-Effectiveness Analysis	Duplicate study
37	Duarte et al. (2014)	Computerised cognitive behaviour therapy for depression management: A cost-effectiveness analysis	Conference abstract
38	Dunlap et al. (2019)	Screening and Intervention for Suicide Prevention: A Cost-Effectiveness Analysis of the ED-SAFE Interventions	Wrong outcomes
39	Ebert et al. (2018)	A health economic outcome evaluation of an internet-based mobile-supported stress management intervention for employees	Wrong patient population
40	Egede et al. (2018)	Cost-Effectiveness of Behavioral Activation for Depression in Older Adult Veterans: In-Person Care Versus Telehealth	Wrong outcomes
41	El Alaoui et al. (2017)	Does internet-based cognitive behaviour therapy reduce healthcare costs and resource use in treatment of social anxiety disorder? A cost-minimisation analysis conducted alongside a randomised controlled trial	Wrong outcomes
42	Esmaeili et al. (2020)	Budget Impact Analysis of a Computer-Delivered Brief Alcohol Intervention in Veterans Affairs (VA) Liver Clinics: A Randomized Controlled Trial	Wrong outcomes
43	Everitt et al. (2019)	Therapist telephone-delivered CBT and web-based CBT compared with treatment as usual in refractory irritable bowel syndrome: the ACTIB three-arm RCT	Wrong patient population
44	Fabian et al. (2017)	Cost-effectiveness of Therapist-guided Internet-delivered Cognitive Behavior Therapy for Pediatric Obsessive-Compulsive Disorder	Duplicate study
45	Fortney et al. (2011)	A budget impact analysis of telemedicine-based collaborative care for depression	Wrong outcomes
46	Garrido et al. (2017)	Computer-assisted cognitive remediation therapy in schizophrenia: Durability of the effects and cost-utility analysis	Wrong outcomes
47	Gerhards et al. (2011)	Economic evaluation of online computerized cognitive behavioural therapy without support for depression in primary care: A randomized trial	Conference abstract
48	Gidding et al. (2018)	PsyScan e-tool to support diagnosis and management of psychological problems in general practice: a randomised controlled trial	Wrong outcomes
49	Godfrey et al. (2005)	Cost effectiveness of treatment for alcohol problems: Findings of the randomised UK alcohol treatment trial (UKATT)	Wrong intervention
50	Grafe et al. (2017)	Internet based treatment of depressive symptoms-a health economic evaluation of costs and benefits	Conference abstract
51	Gräfe et al. (2019)	Health economic evaluation of a web-based intervention for depression: the EVIDENT-trial, a randomized controlled study	Wrong outcomes
52	Gräfe et al. (2020)	Health economic evaluation of an internet intervention for depression (deprexis), a randomized controlled trial	Wrong outcomes
53	Gryczynski et al. (2021)	Computer- vs. nurse practitioner-delivered brief intervention for adolescent marijuana, alcohol, and sex risk behaviors in school-based health centers	Wrong outcomes
54	Hange et al. (2017)	The impact of internet-based cognitive behavior therapy on work ability in patients with depression - a randomized controlled study	Wrong outcomes
55	Havard et al. (2012)	Randomized Controlled Trial of Mailed Personalized Feedback for Problem Drinkers in the Emergency Department: the Short-Term Impact	Wrong intervention
56	Hedman et al. (2011)	Cost-effectiveness of Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: results from a randomized controlled trial	Wrong outcomes

57	Hedman et al. (2013)	Cost-effectiveness and long-term effectiveness of internet-based cognitive behaviour therapy for severe health anxiety	Wrong outcomes
58	Hedman et al. (2014)	Clinical effectiveness and cost-effectiveness of Internet- vs. group-based cognitive behavior therapy for social anxiety disorder: 4-year follow-up of a randomized trial	Wrong outcomes
59	Hedman et al. (2014)	Clinical effectiveness and cost-effectiveness of Internet- vs. group-based cognitive behavior therapy for social anxiety disorder: 4-year follow-up of a randomized trial	Duplicate study
60	Hedman et al. (2016)	Cost effectiveness of internet-based cognitive behaviour therapy and behavioural stress management for severe health anxiety	Wrong outcomes
61	Hedman-Lagerlof et al. (2019)	Cost-Effectiveness and Cost-Utility of Internet-Delivered Exposure Therapy for Fibromyalgia: Results From a Randomized, Controlled Trial	Wrong patient population
62	Henderson et al. (2013)	Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial	Wrong patient population
63	Hollinghurst et al. (2009)	Effectiveness and Cost-Effectiveness of an Internet Based Cognitive Behavioural Psychotherapy for Depression: A Randomised Controlled Trial	Conference abstract
64	Hudson et al. (2017)	Tailored online cognitive behavioural therapy with or without therapist support calls to target psychological distress in adults receiving haemodialysis: A feasibility randomised controlled trial	Wrong study design
65	Isetta et al. (2015)	A bayesian cost-effectiveness analysis of a telemedicine-based strategy for the management of sleep apnea: A multicenter randomized controlled trial	Wrong outcomes
66	Jahoda et al. (2017)	Comparison of behavioural activation with guided self-help for treatment of depression in adults with intellectual disabilities: a randomised controlled trial	Wrong intervention
67	Jahoda et al. (2018)	Behavioural activation versus guided self-help for depression in adults with learning disabilities: the BeatIt RCT	Wrong intervention
68	Kafali et al. (2014)	Cost-effectiveness of a randomized trial to treat depression among Latinos	Wrong outcomes
69	Kaldo et al. (2008)	Internet versus group cognitive-behavioral treatment of distress associated with tinnitus: a randomized controlled trial	Wrong outcomes
70	Kamat et al. (2019)	Effect of video-assisted patient education on compliance with therapy, quality of life, psychomorbidity, and cost of illness in irritable bowel syndrome	Wrong patient population
71	Kemmeren et al. (2016)	The cost-effectiveness of blended cognitive therapy for depression, the e-compared study in the Netherlands	Study protocol
72	Kiluk et al. (2016)	Randomized Trial of Computerized Cognitive Behavioral Therapy for Alcohol Use Disorders: Efficacy as a Virtual Stand-Alone and Treatment Add-On Compared with Standard Outpatient Treatment	Wrong outcomes
73	Klein et al. (2018)	Economic Evaluation of an Internet-Based Preventive Cognitive Therapy With Minimal Therapist Support for Recurrent Depression: Randomized Controlled Trial	Wrong patient population
74	König et al. (2018)	Economic evaluation of cognitive behavioral therapy and Internet-based guided self-help for binge-eating disorder	Wrong comparator

75	Kooistra et al. (2019)	Cost and Effectiveness of Blended Versus Standard Cognitive Behavioral Therapy for Outpatients With Depression in Routine Specialized Mental Health Care: Pilot Randomized Controlled Trial	Wrong comparator
76	Kruger et al. (2014)	The cost-effectiveness of a theory-based online health behaviour intervention for new university students: an economic evaluation	Wrong patient population
77	Lalouni et al. (2018)	Clinical and Cost Effectiveness of Online Cognitive Behavioral Therapy in Children With Functional Abdominal Pain Disorders	Wrong patient population
78	Lavelle et al. (2018)	Cost-effectiveness of collaborative care for depression and PTSD in military personnel	Wrong intervention
79	Le et al. (2019)	The Cost-Effectiveness of an Internet Intervention to Facilitate Mental Health Help-Seeking by Young Adults: Randomized Controlled Trial	Wrong patient population
80	Lenhard et al. (2016)	Cost-effectiveness of internetdelivered cognitive-behavior therapy for obsessive-compulsive disorder: Results from a randomized controlled trial	Conference abstract
81	Lenhard et al. (2016)	Cost-effectiveness of internet-delivered cognitive behavior therapy for adolescent obsessive-compulsive disorder	Conference abstract
82	Lenhard et al. (2020)	Long-term outcomes of therapist-guided Internet-delivered cognitive behavior therapy for pediatric obsessive-compulsive disorder	Wrong outcomes
83	Littlewood et al. (2015)	A randomised controlled trial of computerised cognitive behaviour therapy for the treatment of depression in primary care: the Randomised Evaluation of the Effectiveness and Acceptability of Computerised Therapy (REEACT) trial	Duplicate study
84	Liu et al. (2003)	Cost-effectiveness of collaborative care for depression in a primary care veteran population	Wrong outcomes
85	Ljotsson et al. (2011)	Acceptability, effectiveness, and cost-effectiveness of internet-based exposure treatment for irritable bowel syndrome in a clinical sample: a randomized controlled trial	Wrong patient population
86	Lobban et al. (2020)	A web-based, peer-supported self-management intervention to reduce distress in relatives of people with psychosis or bipolar disorder: the REACT RCT	Wrong patient population
87	Lobban et al. (2020)	Clinical effectiveness of a web-based peer-supported self-management intervention for relatives of people with psychosis or bipolar (REACT): online, observer-blind, randomised controlled superiority trial	Wrong outcomes
88	Lokman et al. (2015)	Return-to-work intervention versus care as usual for sick listed employees with common mental disorders: Trial-based economic evaluation shows promise	Conference abstract
89	Lokman et al. (2017)	Complaint-Directed Mini-Interventions for Depressive Complaints: A Randomized Controlled Trial of Unguided Web-Based Self-Help Interventions	Wrong outcomes
90	Mayoral et al. (2017)	Economic Evaluation of a Guided and Unguided Internet-Based CBT Intervention for Major Depression: Results from a Multicentre Three-Armed Randomized Controlled Trial Conducted in Primary Care	Duplicate study
91	McCollister et al. (2016)	Cost-effectiveness analysis of a continuing care intervention for cocaine-dependent adults	Wrong outcomes
92	McCrone et al. (2004)	Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial	Wrong outcomes

93	McCrone et al. (2007)	Cost-effectiveness of computer-aided behaviour therapy for obsessive-compulsive disorder	Wrong outcomes
94	Moayeri et al. (2018)	Cost-utility analysis of telephone-based cognitive behavior therapy in chronic obstructive pulmonary disease (COPD) patients with anxiety and depression comorbidities: an application for willingness to accept concept	Duplicate study
95	Moessner et al. (2014)	Cost-effectiveness of an internet-based aftercare intervention after inpatient treatment in a psychosomatic hospital	Wrong outcomes
96	Mohr et al. (2019)	A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT	Wrong comparator
97	Moradi-Lakeh et al. (2017)	Cost-effectiveness of aftercare services for people with severe mental disorders: an analysis parallel to a randomised controlled clinical trial in Iran	Wrong intervention
98	Moss-Morris et al. (2015)	A pilot randomized controlled trial of the clinical and cost effectiveness of a skype delivered group mindfulness intervention for distressed people with progressive multiple sclerosis	Conference abstract
99	Mouthaan et al. (2011)	Quality of Life and Cost-Effectiveness of a Brief Web-Based Early Intervention to Prevent PTSD in Traumatic Injury Patients	Study protocol
100	Noben et al. (2014)	Comparative cost-effectiveness of two interventions to promote work functioning by targeting mental health complaints among nurses: pragmatic cluster randomised trial	Wrong outcomes
101	Noben et al. (2015)	Comparative cost-effectiveness of two interventions to promote work functioning by targeting mental health complaints among nurses: Pragmatic cluster randomised trial	Conference abstract
102	Nordgren et al. (2014)	Effectiveness and cost-effectiveness of individually tailored Internet-delivered cognitive behavior therapy for anxiety disorders in a primary care population: a randomized controlled trial	Wrong outcomes
103	O'Connell et al. (2017)	Discrete event simulation modelling of long term cost-effectiveness of internet-based blended cognitive behavioural therapy for major depressive disorder: Extrapolati on of the e-compared randomised controlled trial	Conference abstract
104	Olmstead et al. (2010)	Cost-effectiveness of computer-assisted training in cognitive-behavioral therapy as an adjunct to standard care for addiction	Wrong outcomes
105	Olmstead et al. (2019)	Cost-effectiveness of Electronic- and Clinician-Delivered Screening, Brief Intervention, and Referral to Treatment for Women in Reproductive Health Centers	Wrong outcomes
106	Osborne et al. (2019)	Cost-effectiveness of internet-based cognitive-behavioural therapy for obsessive-compulsive disorder	Wrong study design
107	Painter et al. (2015)	Cost-Effectiveness of Collaborative Care for Depression in HIV Clinics	Wrong intervention
108	Painter et al. (2017)	Cost-Effectiveness of Telemedicine-Based Collaborative Care for Posttraumatic Stress Disorder	Wrong intervention
109	Pot-Kolder et al. (2020)	Cost-Effectiveness of Virtual Reality Cognitive Behavioral Therapy for Psychosis: Health-Economic Evaluation Within a Randomized Controlled Trial	Wrong intervention

110	Pyne et al. (2010)	Cost-effectiveness analysis of a rural telemedicine collaborative care intervention for depression	Wrong intervention
111	Pyne et al. (2015)	Cost-effectiveness of on-site versus off-site collaborative care for depression in rural FQHCs	Wrong intervention
112	Richards et al. (2016)	Clinical effectiveness and cost-effectiveness of collaborative care for depression in UK primary care (CADET): a cluster randomised controlled trial	Wrong intervention
113	Rollman et al. (2012)	The 12-month cost-effectiveness of telephone delivered collaborative care for post-CABG depression	Conference abstract
114	Ruskin et al. (2004)	Treatment outcomes in depression: comparison of remote treatment through telepsychiatry to in-person treatment	Wrong outcomes
115	Salisbury et al. (2017)	An evidence-based approach to the use of telehealth in long-term health conditions: development of an intervention and evaluation through pragmatic randomised controlled trials in patients with depression or raised cardiovascular risk	Wrong intervention
116	Schotanus-Dijkstra et al. (2018)	Towards sustainable mental health promotion: trial-based health-economic evaluation of a positive psychology intervention versus usual care	Wrong outcomes
117	Schubert et al. (2015)	Cost-effectiveness Analysis of a Telephone-based Managed Care Program for Mental Disorders from the Perspective of a Statutory Health Insurance. [German]	No English
118	Sembi et al. (2015)	Mums 4 Mums: Pilot randomised controlled trial of the clinical and cost-effectiveness of telephone peer support for postnatal depression	Conference abstract
119	Shepard et al. (2016)	Telephone-based continuing care counseling in substance abuse treatment: Economic analysis of a randomized trial	Wrong outcomes
120	Simon et al. (2001)	Cost-effectiveness of systematic depression treatment for high utilizers of general medical care	Wrong intervention
121	Simon et al. (2002)	Cost-effectiveness of a program to prevent depression relapse in primary care	Wrong outcomes
122	Simon et al. (2006)	Long-term effectiveness and cost of a systematic care program for bipolar disorder	Wrong outcomes
123	Simon et al. (2009)	Incremental benefit and cost of telephone care management and telephone psychotherapy for depression in primary care	Wrong outcomes
124	Smit et al. (2006)	Cost-effectiveness of preventing depression in primary care patients - Randomised trial	Wrong outcomes
125	Smit et al. (2013)	Cost-effectiveness and cost-utility of Internet-based computer tailoring for smoking cessation	Wrong patient population
126	Solomon et al. (2015)	e-CBT (myCompass), Antidepressant Medication, and Face-to-Face Psychological Treatment for Depression in Australia: A Cost-Effectiveness Comparison	Wrong study design
127	Spindler et al. (2010)	Telehealth in the parkinson's disease subspecialty clinic: The key to the patient-centered medical home	Conference abstract
128	Thase et al. (2020)	Improving Cost-effectiveness and Access to Cognitive Behavior Therapy for Depression: Providing Remote-Ready, Computer-Assisted Psychotherapy in Times of Crisis and Beyond	Wrong comparator
129	Thiart et al. (2016)	Internet-Based Cognitive Behavioral Therapy for Insomnia: A Health Economic Evaluation	Wrong outcomes
130	Titov et al. (2009)	Shyness programme: longer term benefits, cost-effectiveness, and acceptability	Wrong outcomes
131	Valimaki et al. (2017)	Short text messages to encourage adherence to medication and follow-up for people with psychosis (mobile.net): Randomized controlled trial in Finland	Wrong outcomes

132	Van Eeden et al. (2015)	An economic evaluation of an augmented cognitive behavioural intervention vs. computerized cognitive training for post-stroke depressive symptoms	Wrong intervention
133	Van Nispen et al. (2016)	Cost-effectiveness of stepped-care implemented in low vision rehabilitation to reduce depression and anxiety in vision impaired older adults	Wrong intervention
134	Van Spijker et al. (2012)	Reducing suicidal ideation: cost-effectiveness analysis of a randomized controlled trial of unguided web-based self-help	Wrong outcomes
135	Van Spijker et al. (2016)	Online self-help for persons with suicidal intentions: budget impact analysis	Wrong study design
136	Verdonck-De Leeuw et al. (2013)	Cost-evaluation of online guided self-help targeting psychological distress in cancer survivors	Conference abstract
137	Verdonck-De Leeuw et al. (2013)	Efficacy and cost-evaluation of web-based guided self-help targeting psychological distress in cancer survivors	Conference abstract
138	Watson et al. (2018)	Cost-Effectiveness of Internet-Based Cognitive-Behavioral Treatment for Bulimia Nervosa: Results of a Randomized Controlled Trial	Wrong comparator
139	Wijnen et al. (2018)	Complaint-Directed Mini-Interventions for Depressive Symptoms: A Health Economic Evaluation of Unguided Web-Based Self-Help Interventions Based on a Randomized Controlled Trial	Wrong outcomes
140	Wright et al. (2017)	Computerised cognitive-behavioural therapy for depression in adolescents: Feasibility results and 4-month outcomes of a UK randomised controlled trial	Duplicate study
141	Zhou et al. (2019)	Efficacy and cost-effectiveness of internet-based cognitive behavioral therapy for obsessive-compulsive disorder. [Chinese]	No English

Note. Wrong outcomes can mean that no QALYs and/or costs were reported, that QALYs were reported but calculated inadequately or that included costs were not sufficient/appropriate.

C.7 Results on moderator analyses

#	Subgroups	N	Cochrane Q	I^2	Pooled INB	95% CI Pooled INB	P-value
1	Health care perspective	15	$Q(14)=14.2$, $P=.43$	15.1% (0.0%; 64.0%)	$280	$109; $451	.001
	Societal perspective	22	$Q(21)=22.5$, $P=.37$	0.0% (0.0%; 0.01%)	$161	$-247; $569	.44
2	Shorter than 12-month follow-up	14	$Q(13)=5.6$, $P=.96$	0.0% (0.0%; 0.03%)	$112	$-194; $418	.47
	12-month follow-up or longer	23	$Q(22)=30.38$, $P=.11$	12.3% (0.0%; 58.3%)	$270	$-14; $554	.063
3	Depression	16	$Q(15)=13.6$, $P=.55$	0.0% (0.0%; 0.03%)	$387	$156; $618	.001
	Anxiety	7	$Q(6)=1.2$, $P=.98$	0.0% (0.0%; 0.0%)	$644	$227; $1062	.002
	Alcohol or substance abuse	5	$Q(4)=3.6$, $P=.46$	11.6% (0.0%; 56.6%)	$-129	$-448; $191	.43
	Depression and anxiety simultaneously	5	$Q(4)=6.7$, $P=.15$	13.2% (0.0%; 60.3%)	$580	$-584; $1744	.33
	obsessive compulsive disorder	2	$Q(1)=0.2$, $P=.68$	0.0% (0.0%; 0.03%)	$253	$-544; $1051	.53
4	Self-guided intervention	10	$Q(9)=23.7$, $P=.005$	45.2% (36.7%; 89.2%)	$169	$-266; $604	.45
	Guided intervention	27	$Q(26)=13.1$, $P=.98$	0.0% (0.0%; 0.0%)	$317	$84; $550	.008
5	Self-guided intervention	10	$Q(9)=23.7$, $P=.005$	45.2% (36.7%; 89.2%)	$169	$-266; $604	.45
	Less than weekly guidance	3	$Q(2)=0.7$, $P=.71$	0.0% (0.0%; 0.02%)	$108	$-618; $835	.77
	Weekly guidance	21	$Q(20)=9.9$, $P=.97$	0.0% (0.0%; 0.02%)	$413	$146; $680	.002
	More than weekly guidance	3	$Q(2)=0.2$, $P=.90$	0.0% (0.0%; 0.02%)	$-67	$-699; $565	.84
6	Asynchronous guidance	11	$Q(10)=2.4$, $P=.99$	0.0% (0.0%; 0.0%)	$375	$-229; $979	.22
	Synchronous guidance	11	$Q(10)=8.1$, $P=.62$	0.0% (0.0%; 0.02%)	$94	$-335; $524	.67
	Combination	5	$Q(4)=1.1$, $P=.89$	0.0% (0.0%; 0.03%)	$418	$106; $730	.009
7	Open/mass media recruitment	13	$Q(12)=6.2$, $P=.91$	6.9% (0.0%; 16.7%)	$397	$173; $621	.001
	Recruitment by clinical referral	20	$Q(19)=23.7$, $P=.21$	1.7% (0.0%; 14.8%)	$138	$-170; $446	.38
	Other*	4	$Q(3)=4.4$, $P=.22$	0.0% (0.0%; 0.01%)	$91	$-1241; $1423	.89
8	Formal diagnosis for inclusion	13	$Q(12)=11.5$, $P=.48$	0.0% (0.0%; 0.01%)	$311	$-192; $814	.23

	Self-reported symptoms for inclusion	24	$Q(23)=25.5$, $P=.32$	12.4% (0.0%; 58.6%)	$235	$38; $432	.02
9	4-8 weeks intervention duration	18	$Q(17)=10.4$, $P=.89$	0.0% (0.0%; 0.1%)	$398	$209; $587	$<$ .001
	9-12 weeks intervention duration	12	$Q(12)=12.9$, $P=.30$	0.0% (0.0%; 0.02%)	$116	$-313; $546	.60
	Duration more than 12 weeks	3	$Q(2)=0.9$, $P=.65$	0.0% (0.0%; 0.01%)	$-107	$-979; $765	.81
	Undefined intervention duration	4	$Q(3)=6.6$, $P=.09$	25.6% (0.0%; 77.4%)	$-244	$-960; $471	.50
10	Care-as-usual control condition	32	$Q(31)=33.7$, $P=.34$	7.9% (0.0%; 46.1%)	$261	$77; $445	.005
	Attention control condition	5	$Q(4)=3.0$, $P=.55$	0.0% (0.0%; 0.01%)	$64	$-617; $745	.85

CI=confidence interval; INB=incremental net benefit.
Studies falling under this category used a either both open recruitment and clinical referral or screened a specific population.

C.8 Results on sensitivity analyses

#	Subgroups	N	Cochrane Q	I^2	Pooled INB	95% CI Pooled INB	P-value
1	High CHEC list quality rating	24	$Q(23)=18.6$, $P=.72$	0.8% (0.0%; 7.6%)	$253	$43; $463	.018
	Low CHEC list quality rating	13	$Q(12)=18.5$, $P=.10$	26.0% (0.0%; 77.8%)	$197	$-155; $548	.27
2	Low risk of bias rating	8	$Q(7)=3.4$, $P=.84$	0.0% (0.0%; 0.01%)	$244	$-555; $1042	.55
	Medium risk of bias rating	8	$Q(7)=13.0$, $P=.072$	0.0% (0.0%; 0.01%)	$374	$-284; $1031	.27
	High risk of bias rating	21	$Q(20)=20.1$, $P=.45$	9.3% (0.0%; 50.8%)	$287	$113; $461	.001
3	QALY valued at $20,000	37	$Q(36)=36.1$, $P=.46$	5.3% (0.0%; 35.6%)	$145	$56 $234	.001
4	QALY valued at $80,000	37	$Q(36)=40.0$, $P=.30$	12.9% (0.0%; 59.6%)	$431	$115; $747	.008
5	Only studies with directly calculated covariances	12	$Q(11)=6.98$, $P=.80$	0.0% (0.0%; 0.02%)	$264	$-167; $694	.23
	Studies with indirectly calculated covariances	25	$Q(24)=30.1$, $P=.18$	14.0% (0.0%; 61.8%)	$236	$27; $445	.03

CI=confidence interval; INB=incremental net benefit.

Chapter D

Supplemental Material for Chapter 5

D.1 List of all predetermined construct validity hypotheses

Construct validity hypotheses
For all hypotheses we expected a significant medium to high correlation ($0.3 < r < 0.7$) in the direction explained in the article text.

- H1: ICECAP-A capability values and the EQ-5D utility scores;
- H2: ICECAP-A stability subscale and EQ-5D anxiety/depression subscale;
- H3: ICECAP-A attachment subscale and EQ-5D anxiety/depression subscale;
- H4: ICECAP-A autonomy subscale and EQ-5D anxiety/depression subscale;
- H5: ICECAP-A achievement subscale and EQ-5D anxiety/depression subscale;
- H6: ICECAP-A enjoyment subscale and EQ-5D anxiety/depression subscale;
- H7: ICECAP-A autonomy subscale and EQ-5D mobility subscale;
- H8: ICECAP-A autonomy subscale and EQ-5D self-care subscale;
- H9: ICECAP-A autonomy subscale and EQ-5D usual activities subscale;
- H10: ICECAP-A achievement subscale and EQ-5D usual activities subscale;
- H11: ICECAP-A achievement subscale and EQ-5D pain subscale;
- H12: ICECAP-A enjoyment subscale and EQ-5D usual activities subscale;
- H13: ICECAP-A enjoyment subscale and EQ-5D pain subscale.

A hypothesis was added later (not preregistered) to improve the interpretability of the ICECAP-A measurement properties. A strong correlation was expected between the ICECAP-A capability values and a 3-item measure of self-efficacy.

- H14: ICECAP-A capability values and self-efficacy.

Known-group hypotheses
For a hypothesis to be confirmed the differences need to be both statistically significant and greater than the SEM. The SEM can be derived from the error variance of an analysis of variance for repeated measures, including systematic differences: $SEM = \sqrt{\sigma^2_{time} + \sigma^2_{error}}$. Note that hypotheses 19-22 were added later (not preregistered).

- H16: Higher ICECAP-A scores for participants who indicated to be very happy or moderately happy as opposed to participants who indicated to be not very happy or unhappy;
- H17: Higher ICECAP-A scores for participants who indicated to be closer to the best health they could imagine as indicated by the visual analogue scale of the EQ-5D (score between 66 and 100) as opposed to participants indicating being further away from the best health they could imagine (score between 0 and 65). The cutoff score on the visual analogue scale of the EQ-5D for this hypothesis was based on the average and standard deviation of general populations in earlier research, which mostly had a mean of around 80 and standard deviation of around 15 (on a scale of 0 to 100);
- H18: Higher ICECAP-A scores for participants who indicated to have a long-lasting illness as opposed to participants who indicated not to have one;
- H19: Lower ICECAP-A scores for participants who indicated that the long-lasting illness (as reported in H16) obstructed daily life as opposed to participants who indicated that this was not the case;
- H20: Lower ICECAP-A scores for participants who indicated to have been to the hospital in the last three months to visit a doctor as opposed to participants who have not been to the hospital in this period;
- H21: Lower ICECAP-A scores for participants who indicated to have had to stay (spend at least one night) in the hospital in the last three months as opposed to participants for whom this was not the case;
- H22: Lower ICECAP-A scores for participants who indicated to have had at least one visit to the general practitioner in the last three months as opposed to participants for whom this was not the case;
- H23: Higher ICECAP-A scores for people who indicated more self-efficacy in their lives. Self-reported efficacy was assessed with three questions on a 4-point scale (1=often, 2=sometimes, 3=rarely, 4=never) regarding the feeling that one's life is full with possibilities, the feeling to have no control over one's life, and the feeling that one can do the things one wants to do. After recoding the second question, lower scores reflected higher self-reported efficacy. The compared groups were participants who indicated 'often' or 'sometimes' on all three questions versus all other participants;
- H24: Lower ICECAP-A scores for participants who indicated to be unemployed or have an occupational disability as opposed to all other participants;

- H25: Higher ICECAP-A scores for participants who indicated to be in a relationship as opposed to participants who indicated to fall under the category single, divorced, widow or other;
- H26: Higher ICECAP-A scores for participants who indicated to have enjoyed higher education. Three groups were made based on previous research with the EQ-5D [1], being primary and/or lower education, secondary and/or vocational education and higher and/or college education.

References

[1] Janssen, M. F., Pickard, A. S., Golicky, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., & Busschbach, J. (2012). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Quality of Life Research, 22(7), 1717-1727. https://doi.org/10.1007/s11136-012-0322-4

D.2 Comparison of sample with Dutch population on target variables

Variable	Category	Construct validity sample (T1; $N = 941$)	Test-retest sample (T2; $N = 208$)	Dutch population[a]
Age groups	18-24	9.4%	5.8%	11.0%
	25-34	15.0%	9.1%	16.0%
	35-44	14.5%	7.7%	15.0%
	45-54	18.4%	15.9%	18.0%
	55-64	17.6%	21.2%	17.0%
	65-74	20.4%	32.7%	14.0%
	75-99	4.8%	7.7%	10.0%
Gender	Female	51.4%	45.7%	49.11%
	Male	48.4%	54.3%	50.89%
	Other	0.2%	0.0%	Unknown
Region	Groningen	3.5%	5.3%	3.0%
	Friesland	3.4%	5.3%	4.0%
	Drenthe	3.1%	5.8%	3.0%
	Overijssel	8.4%	6.3%	7.0%
	Gelderland	11.6%	11.1%	12.0%
	Flevoland	2.2%	4.8%	2.0%
	Utrecht	8.6%	7.7%	7.0%
	Noord-Holland	16.7%	12.0%	16.0%
	Zuid-Holland	18.5%	16.3%	21.0%
	Zeeland	2.2%	3.4%	2.0%
	Noord-Brabant	14.0%	14.9%	15.0%
	Limburg	7.8%	7.2%	7.0%
Income	<€11.500	5.4%	3.4%	5.0%
	€11.500—€30.000	28.6%	34.6%	26.0%
	€30.000—€36.000	10.4%	11.1%	9.0%
	€36.000—€60.500	31.0%	25.0%	33.0%
	>€60.500	20.7%	22.6%	27.0%
	Rather not tell	3.8%	3.4%	Not applicable
Education	High	37.5%	38.5%	34.2%[b]
	Middle	42.0%	36.5%	37.8%[b]
	Low	20.4%	25.0%	26.3%[b]
	Missing/Unknown	0.1%	0.0%	1.6%[b]

[a] Numbers are based on the latest numbers known to the market research agency unless indicated otherwise.
[b] Numbers are based on 2020 education statistics of the Netherlands' Central Bureau of Statistics.
Note. The selection of a sample representative of the Dutch population was based on the age, gender, region and income variables. Other variables such as education, religion and ethnicity were not considered.

D.3 Individual item details of the ICECAP-A and EQ-5D-5L

Gwet's AC2 [1] was preferred over the intraclass correlation coefficient as test-retest reliability parameter for the individual items of the ICECAP-A and EQ-5D-5L as it is appropriate for ordinal outcomes and skewed data [2, 3]. A Gwet's AC2 of $0.4 - 0.6$, $0.6 - 0.8$ and greater than 0.8 was considered as moderate, good and excellent reliability respectively.

D3.1 ICECAP-A individual item frequencies (%) and reliability for the study sample

Capability	Level 1[a]	Level 2[a]	Level 3[a]	Level 4[a]	Mean (SD)[a]	Gwet's AC2 [95% CI][b]	Level of agreement[b]
Stability	12 (1.3)	108 (11.5)	425 (45.2)	396 (42.1)	3.3 (0.7)	0.64 [0.54; 0.73]	70.7%
Attachment	8 (0.9)	144 (15.3)	382 (40.6)	407 (43.3)	3.3 (0.7)	0.59 [0.49; 0.69]	67.3%
Autonomy	16 (1.7)	79 (8.4)	395 (42.0)	451 (47.9)	3.4 (0.7)	0.62 [0.52; 0.71]	68.8%
Achievement	31 (3.3)	191 (20.3)	456 (48.5)	263 (27.9)	3.0 (0.8)	0.51 [0.39; 0.62]	61.1%
Enjoyment	16 (1.7)	148 (15.7)	422 (44.8)	355 (37.7)	3.2 (0.8)	0.58 [0.48; 0.69]	66.8%

Note. Values represent frequencies with percentages in parentheses unless indicated otherwise. Level 1 corresponds to 'not being able to experience a capability at all' and level 4 to 'being able to fully experience a capability'.
SD = Standard deviation.
[a] Values are based on the total study sample ($N = 941$)
[b] Values are based on the test-retest sample ($N = 208$)

References

[1] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61(Pt 1)*, 29–48. https://doi.org/10.1348/000711006X126600

[2] Tran, D., Dolgun, A., & Demirhan, H. (2020). Weighted inter-rater agreement measures for ordinal outcomes. *Communications in Statistics-Simulation and Computation, 49(4)*, 989–1003. https://doi.org/10.1080/03610918.2018.1490428

[3] Long, D., Polinder, S., Bonsel, G. J., & Haagsma, J. A. (2021). Test–retest reliability of the EQ-5D-5L and the reworded QOLIBRI-OS in the general population of Italy, the Netherlands, and the United Kingdom. *Quality of Life Research*, 1-11. https://doi.org/10.1007/s11136-021-02893-3

D3.2 EQ-5D-5L individual item frequencies (%) and reliability for the study sample

Capability	Level 1[a]	Level 2[a]	Level 3[a]	Level 4[a]	Level 5[a]	Mean (SD)[a]	Gwet's AC2 [95% CI][b]	Level of agreement[b]
Mobility	7 (0.7)	33 (3.5)	66 (7.0)	180 (19.1)	655 (69.6)	4.5 (0.8)	0.75 [0.68; 0.83]	78.4%
Self-care	9 (1.0)	4 (0.4)	18 (1.9)	58 (6.2)	852 (90.5)	4.9 (0.6)	0.92 [0.87; 0.96]	91.8%
Usual activities	12 (1.3)	30 (3.2)	94 (10.0)	201 (21.4)	604 (64.2)	4.4 (0.9)	0.78 [0.71; 0.84]	80.3%
Pain/ discomfort	8 (0.9)	50 (5.3)	131 (13.9)	325 (34.5)	427 (45.4)	4.2 (0.9)	0.59 [0.5; 0.69]	65.9%
Anxiety/ depression	7 (0.7)	28 (3.0)	81 (8.6)	196 (20.8)	629 (66.8)	4.5 (0.8)	0.74 [0.66; 0.81]	76.4%

Note. Values represent frequencies with percentages in parentheses unless indicated otherwise. Level 1 corresponds to 'extreme problems/unable to' and level 5 to 'no problems'.
SD = Standard deviation.
[a] Values are based on the total study sample ($N = 941$)
[b] Values are based on the test-retest sample ($N = 208$)

D.4 Correlation matrix of ICECAP-A and EQ-5D-5L index scores and subscales

	EQ-5D index score	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression	Visual analogue scale
ICECAP capability score	0.60[a]	0.29	0.28	0.50	0.41	0.57	0.58
Stability	0.44	0.13	0.15	0.32	0.30	0.50[a]	0.41
Attachment	0.33	0.11	0.15	0.23	0.16	0.44[a]	0.36
Autonomy	0.45	0.25[a]	0.27[a]	0.44[a]	0.32	0.33[a]	0.42
Achievement	0.53	0.33	0.26	0.48[a]	0.41[a]	0.38[a]	0.51
Enjoyment	0.47	0.24	0.18	0.37[a]	0.34[a]	0.49[a]	0.46

Note. All presented correlations are significant with p-value$< .001$.
[a] Correlation for which predetermined hypotheses were composed.

D.5 Results on hypotheses for known-group differences repeated for the EQ-5D-5L

Hypothesis	Known group	N	Mean rank score	Median	Range	p-value	Confirmed
H16	Happy	800	512	0.9340	$-0.4; 1.0$	$< .001$	Yes
	Unhappy	141	236	0.7540	$-0.1; 1.0$		
H17	VAS $\geq$ 65	714	558	0.9340	$0.0; 1.0$	$< .001$	Yes
	VAS <65	227	197	0.7260	$-0.4; 1.0$		
H18	No illness	562	601	0.9650	$0.3; 1.0$	$< .001$	Yes
	Illness present	379	278	0.7900	$-0.4; 1.0$		
H19a	Non-obstructing illness	51	281	0.9340	$0.3; 1.0$	$< .001$	Yes
	Obstructing illness	328	176	0.7680	$-0.4; 1.0$		
H20	No hospital visit	588	542	0.9340	$-0.1; 1.0$	$< .001$	Yes
	Hospital visit	353	352	0.8340	$-0.4; 1.0$		
H21	No hospital stay	860	485	0.9300	$-0.4; 1.0$	$< .001$	Yes
	Hospital stay	81	319	0.8250	$0.1; 1.0$		
H22	No GP visit	383	582	0.9650	$0.0; 1.0$	$< .001$	Yes
	GP visit	558	395	0.8640	$-0.4; 1.0$		
H23	High self-efficacy	415	583	0.9610	$0.2; 1.0$	$< .001$	Yes
	Low self-efficacy	526	382	0.8640	$-0.4; 1.0$		
H24	Employed	811	504	0.9340	$-0.4; 1.0$	$< .001$	Yes
	Unemployed/ occupational disability	130	265	0.7640	$0.0; 1.0$		
H25	Relationship	640	486	0.9300	$0.0; 1.0$	$= .011$	Yes
	No relationship	301	439	0.8950	$-0.4; 1.0$		
H26b	Higher education	353	NA	0.9340	$0.1; 1.0$	$= .002$	No
	Medium education	395		0.9300	$-0.1; 1.0$		
	Lower education	192		0.8750	$-0.4; 1.0$		

The standard error of measurement (SEM) of the EQ-5D-5L was calculated to be .0133.
GP general practitioner; *VAS* visual analogue scale of the EQ-5D-5L
[a] This question was only applicable to 379 participants who indicated to have a chronic illness
[b] One subject is missing from this analysis since the response to this question was not interpretable

Chapter E

Supplemental Material for Chapter 6

E.1 Orthogonal Main Effects Plan (OMEP) design for the best-worst scaling task

Regular OMEP design						Foldover OMEP design					
Profile	Sta	Att	Aut	Ach	Enj	Profile	Sta	Att	Aut	Ach	Enj
1	2	1	1	2	3	1	3	4	4	3	2
2	1	1	4	3	4	2	4	4	1	2	1
3	4	3	4	2	2	3	1	2	1	3	3
4	1	4	1	4	2	4	4	1	4	1	3
5	1	3	3	1	3	5	4	2	2	4	2
6	4	1	3	4	1	6	1	4	2	1	4
7	2	4	4	1	1	7	3	1	1	4	4
8	3	1	2	1	2	8	2	4	3	4	3
9	1	2	2	2	1	9	4	3	3	3	4
10	2	3	2	4	4	10	3	2	3	1	1
11	3	4	3	2	4	11	2	1	2	3	1
12	3	3	1	3	1	12	2	2	4	2	4
13	2	2	3	3	2	13	3	3	2	2	3
14	4	4	2	3	3	14	1	1	3	2	2
15	3	2	4	4	3	15	2	3	1	1	2
16	4	2	1	1	4	16	1	3	4	4	1

Note. This table shows the levels (ranging from [1] to [4]) on which the corresponding attribute was presented in each of the 16 profiles in the original OMEP design and its foldover.
Sta=Stability, Att=Attachment, Aut=Autonomy, Ach=Achievement, Enj=Enjoyment.

E.2 Demographics, health and ICECAP-A questionnaires

Demographics
Extracted information on demographics was...

1. Age in years;
2. Current living region or province;
3. Gender;
4. Highest completed education level with nine categories (ranging from 'no education' to 'university') that were later transformed to lower, middle and high education;
5. Employment status with eight categories ranging from 'unemployed' to 'retired';
6. Marital status;
7. Household composition.

Health
Extracted information on health was...

1. General happiness on a 4-point scale;
2. General health on a 5-point scale;
3. Chronic illness (yes/no);
4. Whether this illness obstructs daily life in any way (yes/no);
5. The amount of visits to a general practitioner or other doctor;
6. If there were any hospital visits in the last 3 months (yes/no);
7. If there were any hospital stays in the last 3 months (yes/no).

ICECAP-A
The ICECAP-A (Al-Janabi, Flynn & Coast, 2012) measures five capabilities for different aspects of life on a 4-point scale: 1) stability – the extent to which someone can feel settled and secure; 2) attachment – the extent to which someone can feel love, friendship and support; 3) autonomy – the extent to which someone can feel independent; 4) achievement – the extent to which someone can experience achievement and success; 5) enjoyment – the extent to which someone can experience enjoyment and pleasure. Each of the capabilities is presented with four distinct levels (i.e., ranging from [1] not being able to experience a capability at all to [4] being able to fully experience a capability). The five capabilities assess the extent to which someone experiences the freedom to be or carry out what one wishes. Afentou and Kinghorn (2020) have systematically reviewed the literature for studies exploring the psychometric properties of the ICECAP-A. The studies suggest adequate content and construct validity.

References

Al-Janabi H, Flynn TN, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Qual Life Res. 2012;21(1):167-176.

Afentou N, Kinghorn P. A systematic review of the feasibility and psychometric properties of the ICEpop CAPability Measure for Adults and its use so far in economic evaluation. Value Health. 2020;23(4):515-526.

E.3 Best-worst scaling task and accompanying explanations

Participants were randomly assigned; half of the sample to version 1 (OMEP design) and half of the sample to version 2 (OMEP design foldover)

<<Instructions>>
In the next section 16 quality of life situations will be presented to you. The situations will be described based on the five components of quality of life from the previous question. From the five statements you have to choose which statement you find the best (would contribute the most to a valuable life) and which statement you find the worst (would obstruct a valuable life the most).

Here is an example that someone has filled out:

	Best	Worst
I am able to feel settled and secure in a few areas of my life [2]	○	○
I can have quite a lot of love, friendship and support [3]	○	○
I am able to be completely independent [4]	●	○
I can achieve and progress in all aspects of my life [4]	○	○
I cannot have any enjoyment and pleasure [1]	○	●

Just like the previous questions, the number at the end of each statement indicates the level of the statement on a scale of 1 to 4.

Of the above statements this person found 'completely being independent' the best. According to this person, this statement contributes the most to a valuable life, in comparison to the other statements. 'Cannot have any enjoyment and pleasure' has been chosen as worst statement by this person. This statement obstructs a valuable life the most according to this person.

Please read the following statements carefully. Which of the statements do you choose to be the best and which to be the worst? In other words, which statement contributes the most to a valuable life (best) and which statement obstructs having a valuable life the most (worst) according to you?

VERSION 1 (OMEP design)
<<Above each scenario>>
Which statement contributes the most to a valuable life (best) and which the least (worst)?

Scenario 1
I am able to feel settled and secure in a few areas of my life [2]
I cannot have any love, friendship and support [1]
I am unable to be at all independent [1]
I can achieve and progress in a few aspects of my life [2]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 2
I am unable to feel settled and secure in any areas of my life [1]
I cannot have any love, friendship and support [1]
I am able to be completely independent [4]
I can achieve and progress in many aspects of my life [3]
I can have a lot of enjoyment and pleasure [4]

Scenario 3
I am able to feel settled and secure in all areas of my life [4]
I can have quite a lot of love, friendship and support [3]
I am able to be completely independent [4]
I can achieve and progress in a few aspects of my life [2]
I can have a little enjoyment and pleasure [2]

Scenario 4
I am unable to feel settled and secure in any areas of my life [1]
I can have a lot of love, friendship and support [4]
I am unable to be at all independent [1]
I can achieve and progress in all aspects of my life [4]
I can have a little enjoyment and pleasure [2]

Scenario 5
I am unable to feel settled and secure in any areas of my life [1]
I can have quite a lot of love, friendship and support [3]
I am able to be independent in many things [3]
I cannot achieve and progress in any aspects of my life [1]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 6
I am able to feel settled and secure in all areas of my life [4]
I cannot have any love, friendship and support [1]
I am able to be independent in many things [3]
I can achieve and progress in all aspects of my life [4]
I cannot have any enjoyment and pleasure [1]

Scenario 7
I am able to feel settled and secure in a few areas of my life [2]
I can have a lot of love, friendship and support [4]
I am able to be completely independent [4]
I cannot achieve and progress in any aspects of my life [1]
I cannot have any enjoyment and pleasure [1]

Scenario 8
I am able to feel settled and secure in many areas of my life [3]
I cannot have any love, friendship and support [1]
I am able to be independent in a few things [2]
I cannot achieve and progress in any aspects of my life [1]
I can have a little enjoyment and pleasure [2]

Scenario 9
I am unable to feel settled and secure in any areas of my life [1]
I can have a little love, friendship and support [2]
I am able to be independent in a few things [2]
I can achieve and progress in a few aspects of my life [2]
I cannot have any enjoyment and pleasure [1]

Scenario 10
I am able to feel settled and secure in a few areas of my life [2]
I can have quite a lot of love, friendship and support [3]
I am able to be independent in a few things [2]
I can achieve and progress in all aspects of my life [4]
I can have a lot of enjoyment and pleasure [4]

Scenario 11
I am able to feel settled and secure in many areas of my life [3]
I can have a lot of love, friendship and support [4]
I am able to be independent in many things [3]
I can achieve and progress in a few aspects of my life [2]
I can have a lot of enjoyment and pleasure [4]

Scenario 12
I am able to feel settled and secure in many areas of my life [3]
I can have quite a lot of love, friendship and support [3]

I am unable to be at all independent [1]
I can achieve and progress in many aspects of my life [3]
I cannot have any enjoyment and pleasure [1]

Scenario 13
I am able to feel settled and secure in a few areas of my life [2]
I can have a little love, friendship and support [2]
I am able to be independent in many things [3]
I can achieve and progress in many aspects of my life [3]
I can have a little enjoyment and pleasure [2]

Scenario 14
I am able to feel settled and secure in all areas of my life [4]
I can have a lot of love, friendship and support [4]
I am able to be independent in a few things [2]
I can achieve and progress in many aspects of my life [3]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 15
I am able to feel settled and secure in many areas of my life [3]
I can have a little love, friendship and support [2]
I am able to be completely independent [4]
I can achieve and progress in all aspects of my life [4]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 16
I am able to feel settled and secure in all areas of my life [4]
I can have a little love, friendship and support [2]
I am unable to be at all independent [1]
I cannot achieve and progress in any aspects of my life [1]
I can have a lot of enjoyment and pleasure [4]

VERSION 2 (OMEP design foldover)
<<Above each scenario>>
Which statement contributes the most to a valuable life (best) and which the least (worst)?

Scenario 1
I am able to feel settled and secure in many areas of my life [3]
I can have a lot of love, friendship and support [4]
I am able to be completely independent [4]
I can achieve and progress in many aspects of my life [3]
I can have a little enjoyment and pleasure [2]

Scenario 2

I am able to feel settled and secure in all areas of my life [4]
I can have a lot of love, friendship and support [4]
I am unable to be at all independent [1]
I can achieve and progress in a few aspects of my life [2]
I cannot have any enjoyment and pleasure [1]

Scenario 3
I am unable to feel settled and secure in any areas of my life [1]
I can have a little love, friendship and support [2]
I am unable to be at all independent [1]
I can achieve and progress in many aspects of my life [3]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 4
I am able to feel settled and secure in all areas of my life [4]
I cannot have any love, friendship and support [1]
I am able to be completely independent [4]
I cannot achieve and progress in any aspects of my life [1]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 5
I am able to feel settled and secure in all areas of my life [4]
I can have a little love, friendship and support [2]
I am able to be independent in a few things [2]
I can achieve and progress in all aspects of my life [4]
I can have a little enjoyment and pleasure [2]

Scenario 6
I am unable to feel settled and secure in any areas of my life [1]
I can have a lot of love, friendship and support [4]
I am able to be independent in a few things [2]
I cannot achieve and progress in any aspects of my life [1]
I can have a lot of enjoyment and pleasure [4]

Scenario 7
I am able to feel settled and secure in many areas of my life [3]
I cannot have any love, friendship and support [1]
I am unable to be at all independent [1]
I can achieve and progress in all aspects of my life [4]
I can have a lot of enjoyment and pleasure [4]

Scenario 8
I am able to feel settled and secure in a few areas of my life [2]
I can have a lot of love, friendship and support [4]
I am able to be independent in many things [3]

I can achieve and progress in all aspects of my life [4]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 9
I am able to feel settled and secure in all areas of my life [4]
I can have quite a lot of love, friendship and support [3]
I am able to be independent in many things [3]
I can achieve and progress in many aspects of my life [3]
I can have a lot of enjoyment and pleasure [4]

Scenario 10
I am able to feel settled and secure in many areas of my life [3]
I can have a little love, friendship and support [2]
I am able to be independent in many things [3]
I cannot achieve and progress in any aspects of my life [1]
I cannot have any enjoyment and pleasure [1]

Scenario 11
I am able to feel settled and secure in a few areas of my life [2]
I cannot have any love, friendship and support [1]
I am able to be independent in a few things [2]
I can achieve and progress in many aspects of my life [3]
I cannot have any enjoyment and pleasure [1]

Scenario 12
I am able to feel settled and secure in a few areas of my life [2]
I can have a little love, friendship and support [2]
I am able to be completely independent [4]
I can achieve and progress in a few aspects of my life [2]
I can have a lot of enjoyment and pleasure [4]

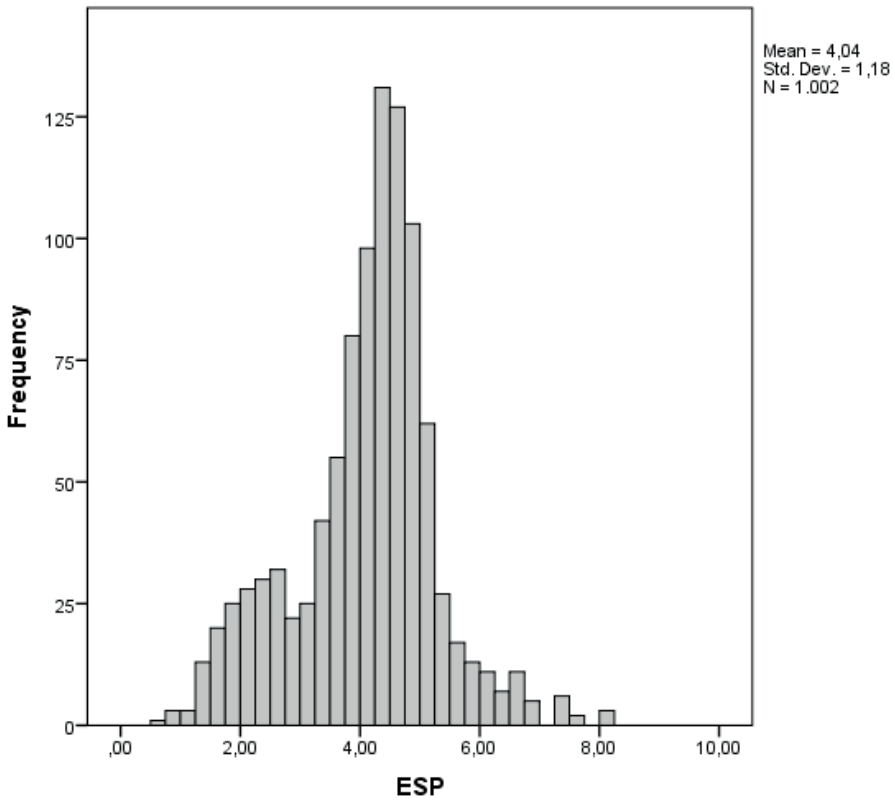
Scenario 13
I am able to feel settled and secure in many areas of my life [3]
I can have quite a lot of love, friendship and support [3]
I am able to be independent in a few things [2]
I can achieve and progress in a few aspects of my life [2]
I can have quite a lot of enjoyment and pleasure [3]

Scenario 14
I am unable to feel settled and secure in any areas of my life [1]
I cannot have any love, friendship and support [1]
I am able to be independent in many things [3]
I can achieve and progress in a few aspects of my life [2]
I can have a little enjoyment and pleasure [2]

Scenario 15
I am able to feel settled and secure in a few areas of my life [2]
I can have quite a lot of love, friendship and support [3]
I am unable to be at all independent [1]
I cannot achieve and progress in any aspects of my life [1]
I can have a little enjoyment and pleasure [2]

Scenario 16
I am unable to feel settled and secure in any areas of my life [1]
I can have quite a lot of love, friendship and support [3]
I am able to be completely independent [4]
I can achieve and progress in all aspects of my life [4]
I cannot have any enjoyment and pleasure [1]

E.4 Distribution of the Empirical Scale Parameter (ESP)



E.5 Comparison of sample with Dutch population on target variables

Variable	Category	Full sample ($N = 1002$)	Analyzed ($N = 933$)	Dutch population[a]
Age groups	18-24	10.6%	10.2%	11.0%
	25-34	15.8%	15.1%	16.0%
	35-44	14.9%	14.9%	15.0%
	45-54	18.4%	18.6%	18.0%
	55-64	16.8%	16.9%	17.0%
	65-74	19.2%	19.6%	14.0%
	75-99	4.5%	4.6%	10.0%
Gender	Female	50.8%	51.3%	49.11%
	Male	49.0%	48.6%	50.89%
	Other	0.2%	0.1%	Unknown
Region	Groningen	3.4%	3.5%	3.0%
	Friesland	3.4%	3.2%	4.0%
	Drenthe	3.3%	3.3%	3.0%
	Overijssel	8.3%	8.6%	7.0%
	Gelderland	11.5%	11.7%	12.0%
	Flevoland	2.3%	2.3%	2.0%
	Utrecht	8.6%	8.6%	7.0%
	Noord-Holland	16.8%	16.7%	16.0%
	Zuid-Holland	18.9%	19.3%	21.0%
	Zeeland	2.2%	2.0%	2.0%
	Noord-Brabant	16.8%	13.3%	15.0%
	Limburg	7.7%	7.5%	7.0%
Income	<€11.500	5.5%	4.8%	5.0%
	€11.500—€30.000	28.2%	28.9%	26.0%
	€30.000—€36.000	10.6%	10.1%	9.0%
	€36.000—€60.500	30.6%	30.9%	33.0%
	>€60.500	21.1%	21.3%	27.0%
	Rather not tell	4.0%	4.0%	Not applicable
Education	High	37.1%	37.6%	34.2%[b]
	Middle	41.7%	42.2%	37.8%[b]
	Low	20.5%	19.5%	26.3%[b]
	Missing/Unknown	0.7%	0.6%	1.6%[b]

[a] Numbers are based on the latest numbers known to the market research agency unless indicated otherwise.
[b] Numbers are based on 2020 education statistics of the Netherlands' Central Bureau of Statistics.
Note. The selection of a sample representative of the Dutch population was based on the age, gender, region and income variables. Other variables such as education, religion and ethnicity were not considered.

E.6 Data quality and exclusion of participants

Excluding individuals from preference studies is not recommended and might lead to selection bias (Lancsar and Louviere, 2006), as it is often unclear whether preference choices are indeed irrational. However, since the best-worst scaling task was conducted anonymous and online and participants received a small monetary sum, some might abuse the assessment. The empirical scale parameter (ESP) is a good tool for detecting such 'gamers'. Indeed, people with an ESP two standard deviations above the average (n=29) showed highly suspicious answering patterns (e.g., always choosing a certain capability as best and another as worst, regardless of the level on which they were presented) and had a significantly lower survey completion time ($M = 9.1$ minutes) compared to the analyzed sample ($M = 14.2$ minutes). Visual inspection of answers from participants with an ESP of two standard deviations below the average ($N = 40$) also revealed unlikely best-worst pairs, suggesting that the task was not understood correctly or taken seriously. Additionally, these participants had a completion time ($M = 9.4$ minutes) that was very similar to participants with a high ESP. Taken together, participants with both an unusually high and low ESP seemed to share characteristics that jeopardized the quality of the research, validating the choice of excluding these participants. Importantly, excluding these participants did not influence the representativeness of the sample (see E.5) or the balance of randomization to the two versions of the best-worst scaling task.

To further explore the impact of exclusion on the quality of the data, best-worst pairs tables were made for different subgroups of the total sample. This made it possible to compare survey completion times and the frequency of unlikely best and worst scores. Table E6.1 lists the findings. We infer from this table that excluding participants with a deviating ESP results in a small improvement in data quality while still retaining a large sample representative of the Dutch population.

References

Lancsar, E., & Louviere, J. (2006). Deleting 'irrational'responses from discrete choice experiments: a case of investigating or imposing preferences? *Health economics, 15(8)*, 797-811.

E6.1 Results on unlikely choices and survey completion time for different sample subgroups

Exclusion criterion	N included	N excluded	Worst choice capability at level 4 in %[a]	Best choice capability at level 1 in %[b]	Survey completion time of excluded participants (SD)
None	1002	0	13.01	6.12	NA
Two SD below or above average ESP	933	69	12.33	5.25	9.3 (6.9)
Two SD below average ESP	962	40	12.72	5.41	9.4 (8.5)
Two SD above average ESP	973	29	12.65	5.99	9.1 (3.9)
Less than 5 minutes completion time	941	61	12.53	5.38	4.4 (0.4)
Two SD below or above average ESP OR less than 5 minutes completion time	881	121	11.87	4.57	7.2 (5.7)

ESP=Emperical scale parameter, SD=Standard deviation

[a] The relative frequency of participants choosing a capability presented at level 4 as worst.

[b] The relative frequency of participants choosing a capability presented at level 1 as best.

E.7 Summary statistics for all estimated SALC models

Preference classes	Scale classes	Worst choice as scale predictor	LL	BIC (LL)	Npar	L^2	Df	p-value	R^2
1-class	-	-	-37922	75973	19	64376	914	$< .001$	0.14
2-class	-	-	-35450	71166	39	59433	894	$< .001$	0.19
3-class	-	-	-34736	69876	59	58005	874	$< .001$	0.23
4-class	-	-	-34342	69224	79	57216	854	$< .001$	0.25
5-class	-	-	-34023	68722	99	56578	834	$< .001$	0.25
1-class	-	yes	-37813	75762	20	64158	913	$< .001$	0.15
2-class	-	yes	-35390	71053	40	59313	893	$< .001$	0.19
3-class	-	yes	-34691	69792	60	57915	873	$< .001$	0.23
1-class	2	-	-35724	71592	21	59981	912	$< .001$	0.19
2-class	2	-	-34804	69889	41	58141	892	$< .001$	0.22
3-class	2	-	-34354	69124	61	57240	872	$< .001$	0.24
1-class	2	yes	-35659	71469	22	59851	911	$< .001$	0.19
2-class	2	yes	-34749	69785	42	58031	891	$< .001$	0.22
3-class*	2	yes	-34284	68992	62	57101	871	$< .001$	0.25
3-class	3	yes	-34316	69069	64	57165	869	$< .001$	0.25

* This model was considered optimal in the current study.

E.8 Attribute importance for the 3 preference classes in the final model

	Class 1	Class 2	Class 3	Total*
Size	0.40	0.30	0.30	1.00
Stability	0.23	0.25	0.21	0.22
Attachment	0.20	0.29	0.30	0.24
Autonomy	0.21	0.24	0.14	0.19
Achievement	0.17	.02	.09	0.13
Enjoyment	0.20	0.20	0.27	0.22

* Attribute importance is weighted by group size and based on the parameters of table 4 in the article text.

E.9 Explorative analyses of subgroup preferences

Preferences for items of the ICECAP-A (e.g., the 'stability' item adds more to well-being than the 'achievement' item) differ not only between countries, but also between groups of people (e.g., older people value 'stability' and 'achievement' differently than younger people). A tariff for a certain population should reflect both these between-item and between-group differences. An additional advantage of exploring subgroups of people who differ in which capabilities they value over others, is that it can be helpful to learn what aspects of quality of life are important for different people. This might eventually translate into more personalized interventions, where the focus of an intervention is adjusted to the values of the patient.

Subgroups were derived from various sociodemographic variables after developing the tariff. First, the sum of squares (based on best-minus-worst scores) of the five capabilities were compared for these subgroups using multiple ANOVA analyses, to separately investigate their relation to capability preferences (see Table E9.1). Second, demographic variables were added to the scale-adjusted latent class analysis to investigate their influence on class membership probability, which is a more rigorous approach to assess heterogeneity in preferences. Subgroups were based on age, gender, marital status, the presence of children, education, employment status, happiness and the presence of a chronic illness. Analyses were exploratory in nature, so should be interpreted with caution.

Subgroup results based on best-minus-worst sum-of-squares comparisons
Exploring preference differences between sum of squares of best-minus-worst scores across different subgroups yielded several results.

1. Higher age was associated with a stronger preference for stability and autonomy and weaker preference for attachment. People aged 40 years or younger found achievement and enjoyment less important than their older counterparts and people aged 41-60 years found enjoyment more important than both older and younger people;
2. Women in the sample valued attachment more than men;
3. People who were in a relationship had a stronger preference for enjoyment than people who were single, divorced or widowed;
4. People without children tended to have a stronger preference for attachment than people with children;
5. There were no apparent differences in preferences between people with different educational levels;
6. Stability was more important for people who indicated to be happy (i.e., very happy or fairly happy) compared to those indicating to be unhappy (i.e., not very happy or unhappy);
7. Preferences did not seem to differ between people with or without a chronic illness;
8. Comparing employed people with people who were unemployed or had an occupational disability yielded no differences in capability preferences.

Subgroup results based on class membership probabilities
Secondly, the demographic variables were added to the final model used for the tariff (see

E9.1 Results of subgroup preferences based on sum of squares comparisons

Group comparisons	N	Mean sum of squares (standard deviation)[a]				
		Stability	Attachment	Autonomy	Achievement	Enjoyment
Age$<= 40$	317	0.92 (0.76)	1.11 (0.81)	0.58 (0.53)*	0.53 (0.50)*	0.70 (0.65)*
$40 <$Age$<= 60$	324	0.80 (0.73)	1.01 (0.84)	0.72 (0.65)*	0.65 (0.66)	0.97 (0.85)*
Age> 60	292	0.99 (0.71)*	0.83 (0.74)*	0.91 (0.78)*	0.67 (0.68)	0.83 (0.66)*
Male	453	0.90 (0.73)	0.87 (0.74)*	0.75 (0.67)	0.58 (0.55)	0.84 (0.76)
Female[b]	479	0.90 (0.75)	1.10 (0.85)*	0.72 (0.67)	0.64 (0.68)	0.83 (0.71)
Relationship	636	0.88 (0.74)	1.01 (0.81)	0.71 (0.64)	0.60 (0.60)	0.88 (0.77)*
No relationship	297	0.94 (0.74)	0.94 (0.80)	0.78 (0.73)	0.65 (0.66)	0.74 (0.66)*
Children	581	0.91 (0.75)	0.94 (0.79)*	0.75 (0.68)	0.64 (0.66)	0.86 (0.75)
No children	352	0.88 (0.72)	1.06 (0.83)*	0.70 (0.66)	0.58 (0.55)	0.79 (0.71)
Higher education	351	0.91 (0.74)	0.99 (0.84)	0.75 (0.67)	0.57 (0.54)	0.88 (0.76)
Medium education	394	0.85 (0.71)	1.02 (0.80)	0.73 (0.68)	0.62 (0.62)	0.80 (0.69)
Lower education[c]	182	0.98 (0.80)	0.91 (0.79)	0.70 (0.66)	0.67 (0.75)	0.81 (0.77)
Happy	797	0.92 (0.75)*	0.99 (0.81)	0.73 (0.64)	0.61 (0.61)	0.84 (0.72)
Unhappy	136	0.78 (0.67)*	0.99 (0.81)	0.77 (0.81)	0.65 (0.69)	0.78 (0.79)
No illness	574	0.93 (0.75)	1.01 (0.83)	0.73 (0.68)	0.59 (0.58)	0.84 (0.73)
Illness present	359	0.84 (0.72)	0.95 (0.77)	0.73 (0.66)	0.64 (0.67)	0.83 (0.74)
Employed	811	0.91 (0.74)	0.98 (0.80)	0.73 (0.66)	0.60 (0.61)	0.83 (0.73)
Unemployed/ occupational disability	122	0.83 (0.75)	1.01 (0.85)	0.78 (0.73)	0.70 (0.65)	0.88 (0.79)

[a] The sum of squares of the five capabilities, based on best-minus-worst scores of participants, were compared for the subgroups using multiple ANOVA analyses, to separately investigate their relation to capability preferences. A alpha value of .05 was maintained to test for significance and multiple testing was not accounted for.

[b] One participant indicated gender as other and is not included in this analysis.

[c] One participant is missing from this analysis since the response to this question was not interpretable.

* Indicates significant difference between (at least one of) the other group(s).

table 4 in the text). Parameters can be found in Table E9.2 The presented odds ratios indicate the probability of a subgroup to be in one class compared to another class. To more easily interpret the results we maintained the following descriptions of the three preference classes based on the parameters of the final model.

- People in class 1 value all five capabilities with no clear preference for one or the other;
- Class 2 is signified by a very low preference for achievement with high preferences for the other capabilities, especially attachment;
- People in class 3 value attachment and enjoyment highly and attach less value to autonomy and achievement.

Some findings can be deduced from these analyses.

1. Younger people are less likely to be in class 2, indicating that achievement is less important for them than their older counterparts. At an older age, the value for capabilities seem to even out as people are more likely to be in class 1;
2. Females are more likely than males to be in class 3, suggesting that they especially value attachment and enjoyment (at the expense of autonomy and achievement);
3. People in a relationship seem to value attachment and enjoyment (at the expense of autonomy and achievement) more than people who are not in a relationship;
4. People without children are more likely to be in class 2, indicating that they attach little value to achievement and much value to attachment;
5. People that enjoyed higher education seem to be less likely to be in class 2, suggesting that they value achievement more (at the expense of attachment);
6. Higher happiness seems to be related with more even preferences over the capabilities (higher probability of being in class 1) and more value attached to achievement (lower probability of being in class 2) compared to people with lower happiness ratings;
7. While the differences were not large, people that are unemployed seem to have stronger preferences (mostly regarding attachment and enjoyment), while employed people do no show preferences and regard all capabilities as equally important;
8. People with a chronic illness are more likely to be in class 2 or 3, possibly because they value attachment and enjoyment (at the expense of autonomy and achievement) more than people without a chronic illness.

Conclusion

While these results are tentative, they show that preferences for domains of quality of life can differ substantially for different people. The results from the sum-of-squares comparisons and class membership probabilities align well and might be interesting starting points for further exploration. Consequently, researchers and clinicians can build on this knowledge by attempting to tailor interventions for specific subgroups, for example by developing interventions for people who are unhappy focused on attachment instead of achievement.

E9.2 Results of subgroup preferences based on class membership probability*

Demographic (Wald, p-value)	Group	Class 1 Coef (OR[a])	Class 2 Coef (OR[a])	Class 3 Coef (OR[a])	OR[b] Class 1-2	OR[c] Class 1-3	OR[d] Class 2-3
Age	$<= 40$	Ref.	Ref.	Ref.			
(24.6, $p < .001$)	$41 <$Age< 60	0.26 (1.30)	-0.57 (0.57)	0.31 (1.36)	2.29	0.95	0.41
	> 60	0.45 (1.57)	-0.28 (0.76)	-0.17 (0.84)	2.08	1.86	0.90
Gender	Male	Ref.	Ref.	Ref.			
(17.7, $p < .001$)	Female	-0.23 (0.79)	-0.27 (0.76)	0.5 (1.64)	1.04	0.48	0.47
Relation	No	Ref.	Ref.	Ref.			
(9.8, $p = .007$)	Yes	-0.17 (0.84)	-0.25 (0.78)	0.42 (1.52)	1.08	0.55	0.51
Children	Yes	Ref.	Ref.	Ref.			
(10.1, $p = .007$)	No	-0.25 (0.78)	0.41 (1.50)	-0.16 (0.85)	0.52	0.91	1.76
Education	High	Ref.	Ref.	Ref.			
(21.9, $p < .001$)	Middle	-0.33 (0.72)	0.34 (1.41)	-0.01 (0.99)	0.51	0.73	1.43
	Low	-0.48 (0.62)	0.64 (1.91)	-0.16 (0.85)	0.32	0.72	2.24
Happiness	Unhappy	Ref.	Ref.	Ref.			
(10.7, $p = .005$)	Happy	0.46 (1.59)	-0.46 (0.63)	0.0 (1.00)	2.53	1.59	0.63
Employment (4.7, $p = .098$)	Unemployed/ occupational disability	Ref.	Ref.	Ref.			
	Employed	0.40 (1.49)	-0.24 (0.79)	-0.16 (0.85)	1.89	1.75	0.92
Illness	Yes	Ref.	Ref.	Ref.			
(10.1, $p = .006$)	No	0.34 (1.41)	-0.29 (0.75)	-0.05 (0.95)	1.88	1.49	0.79

* One participant with other gender was excluded from this analysis to make the data software compatible
Coef=coefficient, OR=Odds ratio, Ref.=Reference.

[a] This signifies the odds ratio of a subgroup being in a preference class compared to the reference (e.g., 1.3 times more likely to be in this class than the reference subgroup)

[b] This signifies the odds ratio of being in preference class 1 compared to class 2 (e.g., 2.29 times more likely to be in class 1 than class 2).

[c] This signifies the odds ratio of being in preference class 1 compared to class 3 (e.g., 0.95 times more likely to be in class 1 than class 3).

[d] This signifies the odds ratio of being in preference class 2 compared to class 3 (e.g., 0.41 times more likely to be in class 2 than class 3).

Chapter F

Supplemental Material for Chapter 7

F.1 Quality-of-life instruments information

F1.1 EQ-5D-5L

The primary outcome measure for the economic evaluation was quality-of-life adjusted life years (QALYs) as assessed with the EQ-5D-5L (EuroQol Group, 1990). The self-report questionnaire measures health related quality of life on five dimensions (i.e., mobility, self-care, usual activities, pain/discomfort and anxiety/depression) on five levels ranging from no disability to extreme disability. The EQ-5D-5L also contains a visual analogue scale on which respondents are asked to rate their current health on a scale from 0 (worst health imaginable) to 100 (best health imaginable). The EQ-5D-5L is widely used and has demonstrated adequate psychometric properties (Feng et al., 2021). The five dimensions with five levels sum up to 3125 possible health states. The Dutch tariff (Versteegh et al., 2016) was used to translate each health state to a utility value anchored at 0 (death) and 1 (perfect health). Utility values were calculated into QALYs over the 14 month follow-up period using the area-under-curve method. This means that utility values were multiplied by the time spent in a certain health state (i.e., 8 weeks or 3 months), where transitions between different health states were linearly interpolated.

F1.2 ICECAP-A

The ICECAP-A (Al-Janabi et al., 2012) presents the five capabilities of stability, attachment, autonomy, achievement, and enjoyment on a four-point scale ranging from not at all to fully being able to experience a capability, and measures the extent to which people are able to do the things they wish. Psychometric properties of the ICECAP-A have been found to be adequate (Afentou & Kinghorn, 2020), also for the Dutch translation (Rohrbach, Dingemans, Essers, et al., 2021). The five capabilities with four levels amount to 1024 possible capability states. Similar to the method used to calculate utility values and corresponding QALYs, a capability value anchored at 0 (no capability) and 1 (full capability) was calculated for each participant using the ICECAP-A Dutch tariffs (Rohrbach, Dingemans, Groothuis-Oudshoorn, et al., 2021) over the 14 month study period.

References

Afentou, N., & Kinghorn, P. (2020). A systematic review of the feasibility and psychometric properties of the ICEpop CAPability Measure for Adults and its use so far in economic evaluation. *Value in Health, 23(4)*, 515-526. https://doi.org/10.1016/j.jval.2019.12.010

Al-Janabi, H., Flynn, T. N., & Coast, J. (2012). Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Quality of Life Research, 21(1)*, 167-176. https://doi.org/10.1007/s11136-011-9927-2

EuroQol Group. (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy, 16(e)*, 199-208. https://doi.org/10.1016/0168-8510(90)90421-9

Feng, Y. S., Kohlmann, T., Janssen, M. F., & Buchholz, I. (2021). Psychometric properties of the EQ-5D-5L: a systematic review of the literature. *Quality of Life Research, 30(3)*, 647-673. https://doi.org/10.1007/s11136-020-02688-y

Rohrbach, P. J., Dingemans, A. E., Essers, B. A., Furth, E. F., Spinhoven, P., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., & Van den Akker-Van Marle, M. E. (2021). The ICECAP-A instrument for capabilities: assessment of construct validity and test–retest reliability in a general Dutch population. *Quality of Life Research, 31*, 687–696. https://doi.org/10.1007/s11136-021-02980-5

Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G. M., Van Til, J. A., Essers, B. A., Spinhoven, P., Furth, E. F., & Van den Akker-Van Marle, M. E. (2021). The ICEpop Capability Measure for Adults instrument for capabilities: development of a tariff for the Dutch general population. *Value in Health, 25(1)*, 125-132. https://doi.org/10.1016/j.jval.2021.07.011

Versteegh, M. M., Vermeulen, K. M., Evers, S. M. A. A., de Wit, G. A., Prenger, R., & Stolk, E. A. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health, 19(4)*, 343-352. https://doi.org/10.1016/j.jval.2016.01.003

F.2 Multiple imputation methods

According to the intention-to-treat approach, all participants who completed baseline were included throughout the analyses. Missing data were multiply imputed (Rubin, 1987) using the software program R version 3.5.1. Categorical variables were imputed using (multinomial) logistic regression. For numerical variables predictive mean matching was used (Rubin, 1986; Van Buuren, 2012). Body weight was skewed to the right at each time point. The weight variables were log transformed to variables closer to normal. Subsequently, linear regression was used to impute these transformed variables. Original weight variables were then imputed through passive imputation (Van Buuren & Groothuis-Oudshoorn, 2011) by back-transforming the log transformed weight variables.

For each variable with missing data a specific number of predictors was used for the prediction of the missing values. This number was determined by using a rule of thumb of 15 cases per predictor (Stevens, 2001). Predictors for the missing data were the variables that were most strongly associated with the variable with missing data. The measure of association used between the variable with missing data and the potential predictor was dependent on the scale level (i.e., numerical or categorical) of both variables. Correlation, partial η^2 and Cramér's V were used for situations where both variables were numerical, variables had a different scale level and both variables were categorical respectively. Missing data were imputed 100 times, creating 100 complete versions of the incomplete dataset.

References

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4(1)*, 87-94. https://doi.org/10.1080/07350015.1986.10509497

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Stevens, I. (2001). *Applied multivariate statistics for the social sciences (4th ed.)*. Psychology Press. https://doi.org/10.4324/9781410604491

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC Press. https://doi.org/10.1201/b11826

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67. https://doi.org/10.18637/jss.v045.i03

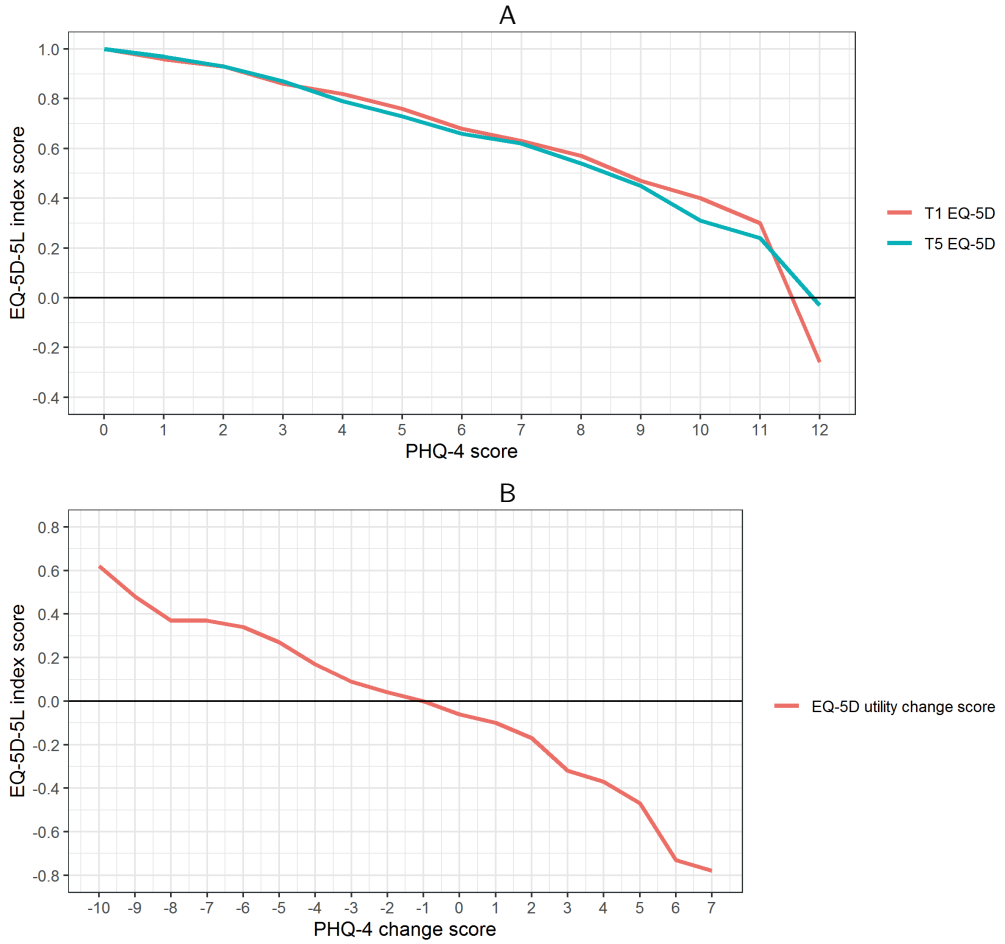
F.3 PHQ-4 to EQ-5D-5L mapping using the equipercentile linking method

In order to map baseline (T0) scores of the PHQ-4 onto the EQ-5D-5L the equipercentile linking method, as explained Kolen and Brennan (2013) and applied by Furukawa et al. (2021), was used. First, Spearman correlations between the PHQ-4 and EQ-5D-5L utility scores were calculated. Spearman correlations were used since the EQ-5D-5L distribution was skewed. Moderate to high correlations (> 0.3) have been successfully used in equipercentile linking. Table F3.1 presents the correlations found between PHQ-4 and EQ-5D-5L scores. Strong correlations were found between the two questionnaires at T1 ($r = -0.66$) and T5 ($r = -0.71$) and a moderate correlation for change scores from T1-T5 ($r = -0.48$). Second, a table with cumulative percentages for the thirteen possible PHQ-4 scores (range 0 – 12) was made for the post-intervention assessment (T1). Corresponding scores of the EQ-5D-5L (based on cumulative percentiles at T1) were identified for the thirteen possible PHQ-4 scores, resulting in a mapping table between the PHQ-4 and EQ-5D-5L. The mapping was applied to PHQ-4 scores at T0 to estimate T0 EQ-5D-5L utility scores. These utility scores were then used for one of the sensitivity analysis described in the main article. Graphical displays of the mapping for various measures of the two questionnaires can be found in Figure F3.2. The final mapping is presented in Table F3.3. Lastly, EQ-5D-5L utility scores for mapped T0, mapped T1 and sample-deduced ('original') T1 scores per study condition can be found in Table F3.4.

F3.1 Spearman correlations between the EQ-5D-5L and PHQ-4 scores at T1, T5 and change scores (T1-T5)

	T1 PHQ	T5 PHQ	T1 EQ-5D	T5 EQ-5D	T1-T5 change PHQ	T1-T5 change EQ-5D
T1 PHQ	1.00	.52	-.66	-.47	-.49	.23
T5 PHQ		1.00	-.51	-.71	.42	-.20
T1 EQ-5D			1.00	.64	.17	-.40
T5 EQ-5D				1.00	-.24	.35
T1-T5 change PHQ					1.00	-.48
T1-T5 change EQ-5D						1.00

F3.2 Equipercentile mapping plots for (A) PHQ-4 and EQ-5D-5L at T1 and T5, and (B) PHQ-4 and EQ-5D-5L T1-T5 change scores



F3.3 Final mapping table based on T1 scores of the PHQ-4 and EQ-5D-5L

PHQ-4 score	EQ-5D-5L score	Cumulative percentile
0	1.00	1.1
1	0.96	3.7
2	0.93	6.8
3	0.86	13.0
4	0.82	22.8
5	0.76	31.5
6	0.68	42.8
7	0.63	54.9
8	0.57	67.6
9	0.47	78.3
10	0.40	84.8
11	0.30	92.7
12	-0.26	100.0

F3.4 Mean EQ-5D-5L utility scores and standard errors using the final mapping

EQ-5D-5L utility	Featback (N=88)	Featback + Expert-patient support ($N = 90$)	Expert-patient support ($N = 87$)	Waiting list ($N = 90$)	Total sample ($N = 355$)
Baseline (T0): mapped	0.49 (0.04)	0.48 (0.04)	0.45 (0.04)	0.43 (0.04)	0.46 (0.02)
Post intervention (T1): mapped	0.62 (0.03)	0.63 (0.03)	0.54 (0.03)	0.49 (0.04)	0.57 (0.02)
Post intervention (T1): original	0.68 (0.03)	0.68 (0.03)	0.61 (0.03)	0.58 (0.03)	0.64 (0.01)

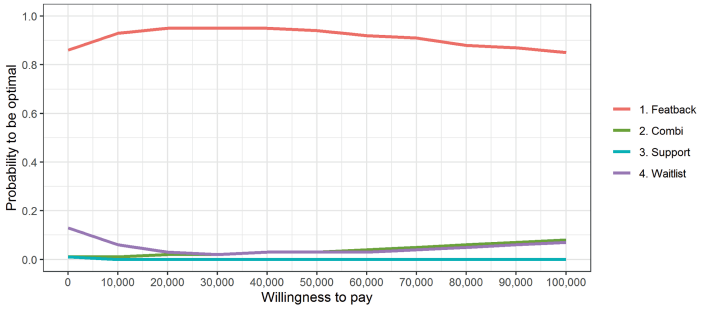
References

Furukawa, T. A., Levine, S. Z., Buntrock, C., Ebert, D. D., Gilbody, S., Brabyn, S., Kessler, D., Björkelund, C., Eriksson, M., Kleiboer, A., Van Straten, A., Riper, H., Montero-Marin, J., Garcia-Campayo, J., Phillips, R., Schneider, J., Cuijpers, P., & Karyotaki, E. (2021). How can we estimate QALYs based on PHQ-9 scores? Equipercentile linking analysis of PHQ-9 and EQ-5D. *Evidence-Based Mental Health, 24(3)*, 97-101. http://dx.doi.org/10.1136/ebmental-2020-300240

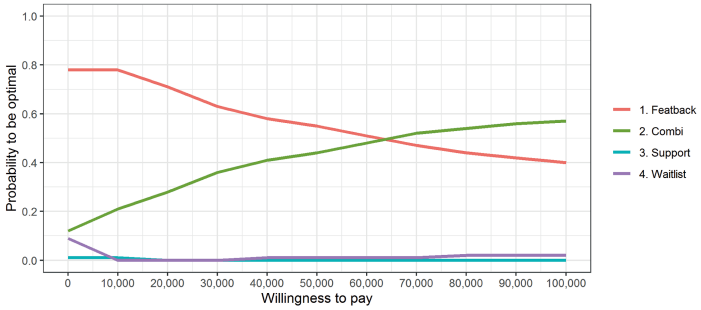
Kolen, M. J., & Brennan, R. L. (2013). Test equating: Methods and practices. Springer Science & Business Media.

F.4 Results of sensitivity analyses

F4.1 Cost-utility acceptability curves with the EQ-5D-5L visual analogue scale



F4.2 Cost-utility acceptability curves with direct health care costs only



F4.3 Cost-utility acceptability curves with baseline EQ-5D-5L values derived from the PHQ-4 equipercentile mapping procedure