A chance for change : building an outcome monitoring feedback system for outpatient mental health care

Jong, K. de

A chance for change: Building an outcome monitoring feedback system for outpatient mental health care

Kim de Jong

A chance for change

Building an outcome monitoring feedback system
for outpatient mental health care

The work presented in this dissertation was performed at GGZ Noord-Holland-Noord, Research and Monitoring department, Heiloo and at the Department of Medical Psychology and Psychotherapy, Erasmus University Medical Center, Rotterdam. Printing this dissertation was financially supported by both.

The studies presented in Chapter 2, 3, 4 and 5 were supported by GGZ Noord-Holland-Noord and the participatng mental health care organizations. The study presented in Chapter 6 in this dissertation was supported by a grant from The Netherlands Organisation for Health Research and Development (ZonMW), grant number 94506414.

Jong, Kim de
A chance for change: Building an outcome monitoring feedback system for outpatient mental health care
Dissertation Leiden University – With summary in Dutch

Subject headings:

Cover art by Sarah Atzori
Layout by Ralph Feenstra
Printed by Ridderprint Grafisch Bedrijf, Ridderkerk

ISBN: 978-90-5335-531-2

A chance for change

Building an outcome monitoring feedback system for outpatient mental health care

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. Mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
ter verdediging op dinsdag 17 april 2012
klokke 13.45 uur

door

Kim de Jong

geboren te Alkmaar in 1977

Promotiecommissie

Promotores:	Prof. dr. P. Spinhoven Prof. dr. W. J. Heiser
Co-promotor:	Dr. M.A. Nugter (GGZ Noord-Holland-Noord)
Overige leden:	Dr. I.V.E. Carlier (Leids Universitair Medisch Centrum) Prof. dr. I.M. Engelhard (Universiteit Utrecht) Prof. dr. W. Lutz (Universität Trier, Deutschland)

Contents

Chapter 1

General Introduction

Background

Over the last decades critique on the generalizability of scientific studies on the outcome of psychological interventions (e.g. Essock, Drake, Frank, & McGuire, 2003; Westen, 2005; Stirman, DeRubeis, Crits-Christoph, & Brody, 2003) has led to the emergence of practice-based research within the scientific community (Margison et al., 2000). At the same time, there was an increasing demand for practice-based research from therapists and mental health care providers. A combination of factors, such as expanding costs of mental health care and a deteriorating economic situation, has led to political measures, such as a restriction of the amount of refundable therapy sessions by insurance companies (e.g. Lambert, Huefner, & Nace, 1997). This state of affairs has made it necessary for providers to show that their interventions are effective. Measuring outcomes was also stimulated by the quality assurance perspective that has been adopted by the field (e.g. Edmunds et al., 1997; Valenstein et al., 2004).

Outcomes of psychological interventions depend on many factors, including patient and therapist characteristics and treatment variables. Results from randomized controlled trials tend to show positive results of treatment, but results measured in clinical practice are less favourable (Barkham et al., 2008; Hansen, Lambert, & Forman, 2002; Hansen, Lambert, & Forman, 2003; Weisz, Donenberg, Han, & Weiss, 1995). There has been a growing awareness in research and practice that treatment effects should be evaluated for the individual patient in everyday practice and that the treatment has to be adjusted when patients do not progress according to expectation (Lambert et al., 2003). Measuring patient progress in clinical practice is often referred to as (routine) outcome monitoring or measuring. In outcome monitoring therapists are typically provided with feedback on their patients' progress using a generic outcome measure. In some models, the progress of the patient is benchmarked against a prediction model (Lambert, 2007). Recently, outcome monitoring has burgeoned and there is a variety of feedback models available, that not only differ in design, but also in effectiveness. Recent review articles and meta-analyses show that feedback can have slightly negative to very large positive effects (Carlier et al., 2010; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Shimokawa, Lambert, & Smart, 2010). In addition to therapists being able to adjust the treatment of individual patients, outcome monitoring enables research on patient, therapist and treatment factors that may be predictive of positive outcomes, which may inform decision making on an aggregated level (e.g. treatment unit, organization, national or international level).

At the start of this project, Michael Lambert was one of the few who had done controlled research in this area. His group still has performed the largest number of studies, but it was uncertain whether results of these studies could be generalized to Dutch outpatients, since his research was carried out in a very specific setting. The

majority of the patients in his studies were students at a private mormon (Latter Day Saints; LDS) university, who were seeking counseling in the university counseling center, resulting in highly educated, young patients of relatively high social economic status (Shimokawa, et al., 2010). In addition, practically all of the therapists are LDS church members as well. This resulted in a very specific cultural subgroup of patients and therapists and the results of his studies might not translate well to other American cultural groups, let alone Dutch patients and therapists. Apart from cultural differences, the American mental health care system is different from the system in the Netherlands, where almost all mental health care is in the public sector and even service provided by private practitioners is reimbursed by public funds.

The central objective of this thesis is to develop an effective outcome monitoring feedback model that is fit for outpatient mental health treatment in the Netherlands and that may improve treatment outcomes for patients. The purpose of this chapter is to give an overview of the relevant literature. The chapter is organized as follows. First, an overview of relevant literature on outcome monitoring is given. Second, a definition of outcome, treatment success and treatment failure is provided. Third, relevant predictors of outcome will be discussed. Finally, the need for feedback to therapists is discussed. The chapter ends with an overview of this thesis.

Outcome monitoring

As has been stated above, outcome monitoring refers to the (frequent) measuring of results during therapy. The measures that are used depend on the treatment goals, which may differ distinctly across treatment settings. Worldwide, there are many large initiatives in which outcomes are monitored in clinical practice, including the COMPASS (Howard, Moras, Brill, Martinovich, & Lutz, 1996), the Partners for Change Outcome Management System (Miller, Duncan, Sorrell, & Brown, 2005), and the Treatment Outcome Package (Kraus, Seligman, & Jordan, 2005) in the United States; the Clinical Outcomes in Routine Evaluation – Outcome Measure (Evans et al., 2002) in the United Kingdom; and the Health of the Nation Outcome Scales (HoNOS) in the United Kingdom, Australia and New Zealand (Burgess, Pirkis, & Coombs, 2006; Wing et al., 1998).

In the Netherlands, a large national project on routine outcome monitoring started in 2009. The project is a collaboration between the branch organization of mental health care providers, the branch organization of health insurance companies, the national patient platform for mental health care, the Dutch institute for psychologists, the Dutch organization for psychiatrists and the national organization for patients in mental health care. It aims to promote routine outcome monitoring in clinical practice for all patient groups (Stuurgroep ROM ggz, 2010). What is unique about the project is that it is not centered around the use of one instrument or patient group, but a wide range

of instruments and patients. There are several task forces in the project that advice the field on topics such as the use of instruments, measurement frequency, comparability of outcome measured by different instruments, implementation and infrastructure.

Outcome monitoring has been promoted by several professional organizations. The American Psychological Association (APA) Presidential Task Force on Evidence-Based Practice in Psychology (2006) states that "providing clinicians with real-time patient feedback to benchmark progress in treatment and clinical support tools to adjust treatment as needed" is one of "the most pressing current research needs" (p.278). The APA Interdivisional (Divisions 12 and 29) Task Force on Evidence-based Therapy Relationships (2011) reports that collecting client feedback is demonstrably effective and states in their recommendations for practice: "Practitioners are encouraged to routinely monitor patients' responses to the therapy relationship and ongoing treatment. Such monitoring leads to increased opportunities to reestablish collaboration, improve the relationship, modify technical strategies, and avoid premature termination". Similar advice is given by the American Group Psychotherapy Association (2007) in their Practice Guidelines for Group Psychotherapy.

Outcome monitoring can have several functions. The ROM ggz project mentions four functions of outcome monitoring: 1) to support clinicians in their treatment process; 2) to learn from aggregated outcomes within the organization and compare oneself with other organizations; 3) to demonstrate the effectiveness of treatments to third-party payers, such as insurance companies and the Ministry of Health; 4) to study predictors and mediators of outcome using large national databases (Stuurgroep ROM ggz, 2010). Kazdin (2008) states that the key argument for systematic evaluation is to provide high-quality care. Whether clinicians use evidence-based treatment or individualized treatment, they can never be sure that the treatment will be effective. In addition, monitoring treatment effects in an ongoing way is important to make decisions on continuing, altering or terminating treatment on the basis of how well the patient is doing (Kazdin, 2008). Sapyta, Riemer and Bickman (2005) use the metaphor of learning archery and indicate that without feedback on where the arrow lands, it is impossible to master archery, no matter how much natural talent someone has at archery. They state that: "Without direct feedback on how their clients are progressing, clinicians are essentially wearing a blindfold while shooting at a target". Although the archery metaphor is certainly appealing, in real life the analogy does not hold. In archery feedback is straightforward: you hit the target or not. In clinical practice, the feedback will not always be that clear. When two measurements are administered to a patient, how does the clinician know if he or she is on target or not? How much does a patient need to improve to be considered a successful case? These questions will be addressed in the next paragraph.

Defining outcomes

Treatment failure

One of the main objectives of feedback models in clinical practice is to prevent treatment failure. Until recently, treatment failure was an issue that was not often addressed in literature, but a renewed interest in the matter has emerged, resulting in special issues in *Cognitive Behavioral Practice* (Dimidjian & Hollon, 2011) and the *Journal of Clinical Psychology* (Lampropoulos, 2011), as well as a series of articles in the *American Psychologist* (Barlow, 2010; Castonguay, Boswell, Constantino, Goldfried, & Hill, 2010; Dimidjian & Hollon, 2010). Treatment failure is a complex concept to define, as it may include dropout or premature termination, nonresponse (no change), partial change, slow change, deterioration (negative change) and relapse (failure to maintain gains). In addition, there might be challenges in defining failure in therapy depending on whose perspective is being used: patient, therapist or other observers (Lampropoulos, 2011). It seems that definition of treatment failure also differs between different treatment modalities. A recent special issue of the *Journal of Clinical Psychology* (November 2011) showed treatment failure from the perspective of five different theoretical orientations. In psychodynamic therapy gaining insight and character change are considered the most important goals and treatment could be perceived as failed by psychodynamic therapists if symptomatic relief would occur without structural changes in character (Gold & Stricker, 2011). The behavioral therapy approach focuses on nonresponse and deterioration, defined by symptom attenuation and uses self-report measures to assesses outcomes (Hopko, Magidson, & Lejuez, 2011). Watson (2011), who is an experiential psychotherapist, stresses that for patients who are nonresponders or who are slow to respond, the question of whether treatment has failed is not as easy to answer. In interpersonal therapy, failure includes partial response, nonresponse, worsening and premature termination (Ravitz, McBride, & Maunder, 2011). Common denominators that capture most perspectives on treatment failure across theoretical orientations include nonresponse and deterioration (Lambert, 2011), which is the definition we use in this thesis.

Clinical significance

Nonresponse and deterioration are most often defined by the concept of clinical significant and reliable change (Jacobson & Truax, 1991). These authors define clinical significance as returning to normal functioning and this approach is nowadays the leading method in outcome research. The criterion is twofold: (a) the magnitude of change has to be statistically reliable and (b) by the end of treatment patients have

to end up in a (score) range that renders them either indistinguishable from well-functioning people, or has them in the lower range of dysfunctioning (Jacobson, Roberts, Berns, & McGlinchey, 1999). A cut-off point for normal functioning and a reliable change index are calculated. The method works best when adequate norms are available for both the dysfunctional and the normal population. Three possible ways to calculate the cut-off point for normal functioning are given. The first method uses a score of two standard deviations above the patient population mean as cutoff. The second method is similar to the first and uses a score of two standard deviations below the normal population mean as a cut-off point. These two methods can only be used when the population curves of the normal population and patient population do not overlap. The third method is the most frequently used method and places the cut-off point at the intersection of the dysfunctional and normal population curves. It is estimated by using the mean and standard deviation of the normative samples for both populations. The reliable change index (RCI) is the minimum amount of change that has to occur to be statically reliable. It is usually expressed as the amount of points on a certain measurement instrument that a patient has to improve between pre- and post-treatment measurements. The RCI depends on the reliability of the measurement instrument and the variability of scores.

Within this system a patient is classified as 'recovered' if reliable change has occurred and the cutoff point for clinical significance is crossed. If the cutoff point is not crossed, but reliable change took place, a patient is classified as 'improved'. If reliable change occurs in opposite direction, the patient is classified as 'deteriorated'. Patients are considered nonresponders if they do not improve nor deteriorate (Jacobson, Follette, & Revenstorf, 1984; Jacobson, et al., 1999; Jacobson & Truax, 1991).

There are some limitations to defining treatment success in this way. If the patients end up in the normal range at the end of treatment, but the magnitude of change is not reliable, patients are considered nonresponders, even though they are functioning comparable to non-patients. In addition, if a patient starts treatment in the normal range, no clinical significance can occur. This is especially problematic in naturalistic studies where selection on severity of problems is less common than in controlled trials. On the other end of the spectrum, there are patients who are not likely to ever return to normal functioning and for whom clinical significant change might not be the best criterion for treatment success. Nevertheless, for the majority of patients in outpatient mental health care it will be possible to classify the outcome of therapy according to this method. Despite its drawbacks, the method of clinical significance by Jacobson and colleagues remains the most frequently applied method and will be applied throughout this thesis.

Outcome: rate of change and end state functioning

The method of determining clinical significance provides us with information on how to interpret the result of treatment when therapy has ended, but does not describe the way in which patients change between the beginning and end of therapy. Recently, there is a renewed interest among psychotherapy researchers in studying not only whether treatments work, but also for whom and under what conditions they work, how they work, and why they work (Laurenceau, Hayes, & Feldman, 2007). Outcome monitoring provides data on the course of therapy and it makes sense to analyse all available data rather than just begin and end state functioning. In fact, one of the major problems in measuring outcomes in psychotherapy research is that it can be a challenge to collect data on treatment outcomes at the end of treatment, due to reasons such as drop-out from therapy or research. This difficulty is especially relevant for practice-based studies, with large numbers of patients and reduced control over the measurements by the researchers.

The availability of large naturalistic data sets that are the result of practice based studies has fired an interest in longitudinal models that are more flexible in handling unbalanced longitudinal data, including survival analysis and multilevel analysis (Singer & Willet, 2003). These models tend to describe change over time and predictor variables are modelled as interaction with time, meaning that they predict the speed of recovery – referred to as the rate of change – rather than end state functioning. To illustrate the difference between outcome defined as end state function and rate of change, consider the following example: two patients can have the exact same score at the start of therapy and at the end of therapy, so the absolute change they made is equal. However, Patient 1 received 10 sessions of therapy, whereas Patient 2 needed 30 sessions. Both have the same end state functioning, but the rate of change is higher in Patient 1. Alternatively, two other patients could have the same starting point and the same rate of change but different treatment durations, resulting in differences in end state functioning. In this thesis, both end state functioning and rate of change will be used as outcome variables.

Predictors of outcome

A vast amount of literature exists on the subject of predicting treatment outcomes. It would be beyond the scope of this thesis to give a full review of all the research that is done in this area, since sufficient review literature is available (e.g. Beutler et al., 2004; Clarkin & Levy, 2004; Lambert & Ogles, 2004; Orlinsky, Ronnestad, & Willutzki, 2004). It is relevant to note that there are two lines of research that look into predictor variables: one focusing on treatment-specific effects and one focusing on common

or non-specific factors. Lutz (2002) states that identifying particular treatments for particular diagnoses could be seen as a bottom-up approach, whereas the top-down approach assesses a class of treatments defined by overlapping techniques, mechanisms and proposed outcomes and could be seen as a first step for large-scale outcome monitoring. Both are relevant for the field of outcome research. For the current research we are interested in the full range of patients that are in outpatient mental health care and not in specific groups of patients. Therefore we take the top-down approach and look for common factors that predict treatment outcomes.

Predictor variables with a demonstrated relationship with treatment outcome can be classified into three categories: patient characteristics, therapist characteristics and treatment characteristics. The aim of the prediction model is to predict the treatment course and end state functioning of new patients; therefore, the focus is on pre- and early treatment predictors. that can be identified at the beginning of treatment, to make early prediction possible. Patient characteristics include both sociodemographic variables (e.g. sex, age, marital status) and clinical variables (e.g. diagnosis, prior treatment). The influence of demographic variables on outcome is mixed and inconsistent (Clarkin & Levy, 2004). Clinical variables have been studied intensively and frequently found predictors include initial severity of symptoms, having multiple Axis I or Axis II diagnoses, or a combination of Axis I and II disorders, prior treatment and duration of psychiatric complaints.

A second category of outcome predictors are characteristics of the treatment received by the patient. These characteristics include a wide range of variables associated with therapeutic techniques, the duration and frequency of therapy sessions, the treatment modality (e.g. individual, group, couples counselling, etc) and interactions between the patient and therapist. It is beyond the scope of this thesis to discuss the full range of treatment predictors. In naturalistic studies, treatment variables usually are outside the control of the researcher and treatments are often modified during the treatment process, making treatment characteristics complicated to use as predictors. Variables that are of interest to us are two early treatment factors that are consistently associated with outcomes: expectancies and the working alliance. Although expectancies could be considered patient characteristics, they usually emerge in the interaction with the therapist and can be modified within that interaction (Westra, Constantino, & Aviram, 2011). Expectancies have a small positive effect on outcome (Constantino, Arnkoff, Glass, Ametrano, & Smith, 2011). The relationship between therapist and patient, referred to as the working alliance or therapeutic alliance, has been the focus of many studies and appears to be a consistent albeit moderate predictor of treatment outcome (Hentschel, 2005; Horvath & Symonds, 1991; Martin, Garske, & Davis, 2000).

Numerous studies have demonstrated that therapist factors explain part of the

variance in outcomes between patients (Crits-Christoph et al., 1991; Dinger, Strack, Leichsenring, Wilmers, & Schauenburg, 2008). Literature shows that therapist effects are usually small in randomized controlled trials and small to medium in naturalistic studies (Crits-Christoph & Gallop, 2006). Little is known about variables that predict therapist effects: demographic characteristics fail to emerge as predictors of outcome, but variables that are more closely related to personality traits for the therapist have been found as predictors, although more research is needed on the subject (Anderson, Ogles, Patterson, Lambert, & Vermeersch, 2009; Beutler, et al., 2004). In order to predict the treatment course and end state functioning for new patients, therapist variables may not be of much use, since it is still quite unclear how they affect outcomes. However, what is interesting for this thesis is how therapist variables may interact with the use of feedback by the therapist. Therefore, therapist variables that have found to be related to acceptance and use of feedback will be included in the study.

Feedback to therapists

The need for feedback

One could wonder why therapists need feedback on their patients' progress. In fact, many therapists claim that they do not need feedback and are perfectly able to monitor progress in their patients themselves. However, meta-analysis comparing 67 studies showed that statistical prediction had somewhat greater accuracy than clinical judgement (Ægisdóttir et al., 2006). In his review of the literature on clinical decision making, Garb (2005) concludes that psychologists should reduce their reliance on informal observation and clinical validation when making clinical judgements or interpret tests. None of these studies have been done on decisions on patient progress. In a study more specifically tailored to recognizing patients that deteriorate, Hannan and colleagues (Hannan et al., 2005) asked therapists to rate their entire caseload on the risk of deterioration for three consecutive sessions. Therapist were able to identify less than 1% of patients that deteriorated. The statistical model, using expected treatment recovery curves, was able to correctly identify 85% of all patients that deteriorated. More recently, Hatfield and colleagues showed that therapists had considerable trouble in identifying deteriorating clients when using their case files. Although therapists did seem to recognize the factors that might be related to deterioration, such as symptom increase and therapy indications such as ruptures in the working alliance or treatment goal failure, they still had trouble noticing these cues in their patients (Hatfield, McCullough, Frantz, & Krieger, 2010).

That clinicians may need feedback on their patients in order to prevent negative outcomes, does not mean that they like it. Some clinicians do not want to use outcome

monitoring feedback, even if it is beneficial to the patient (Aoun, Pennebaker, & Janca, 2002; Walter, Cleary, & Rey, 1998). Feedback theory provides several factors that might be related to the attitude that therapists have towards feedback and whether they will use it. Clinicians will be more likely to use feedback effectively if they perceive the feedback as valid and reliable, pay attention to the feedback, have a preference for externally generated feedback and are committed to use the feedback in their treatment (Claiborn & Goodyear, 2005; Herold & Fedor, 2003; Riemer & Bickman, 2011).

Towards building an outcome monitoring feedback model

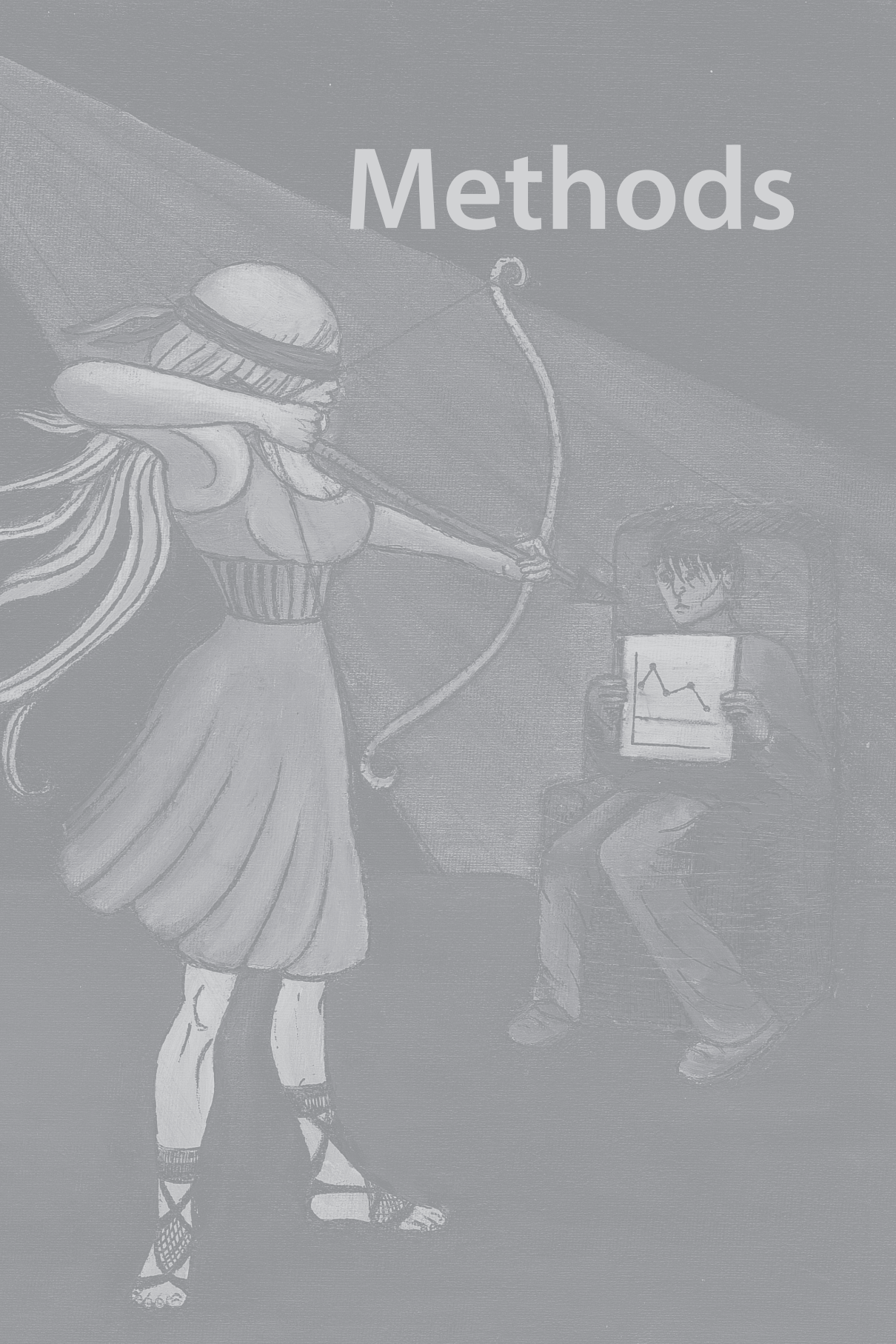
As was mentioned before, the main objective of this thesis was to develop an outcome monitoring feedback model for outpatients in the Netherlands and test whether providing feedback on patient progress to therapists can improve treatment outcomes. Therefore, the first step in the process of building a feedback model was to get an idea on how Dutch patients score on the outcome measure we wanted to use – the Outcome Questionnaire – 45 (OQ-45; Lambert et al., 2004) (see Chapter 2) and to obtain more insight in the progress in therapy that is made by these patients over time. In addition, we were interested in the pre and early treatment variables that predict outcomes and the differences among predictive variables in predicting the rate of change and end state functioning (Chapter 4).

In designing a study to measure the effect of feedback, we needed to take into account that the feedback about the patient's progress is provided to therapists (and sometimes the patients) and therapists might differ in how they handle the feedback, which in turn might influence the effectiveness of the feedback. In the study design, and more specially in estimating the amount of patients that needed to be included in the study to have sufficient statistical power, we needed to take into account that the data have a hierarchical structure with three levels: patients were nested within therapists and measurements were nested within patients. In Chapter 3 the factors that influence study design and study power in a situation with therapists accounting for some of the variance of the outcomes in patients is described.

Finally, two feedback studies were conducted. The first study was a multicenter study in an outpatient mental health care setting and followed patients in their treatment for one year. Feedback was provided to the therapist only and in addition to patient outcomes, and therapist characteristics on relevant traits that are hypothetically related to feedback effectiveness were studied (see Chapter 5). The second study was conducted in both public outpatient centers and in private practices and contained both short and longer term therapies, up until two years. This study had three treatment conditions, a control group, a group with feedback to therapists alone and a group in which both therapists and patients received feedback about the patient's

progress (see Chapter 6). Finally, in the general discussion, results of the studies will be summarized and discussed, study strengths and limitations will be reviewed and implications for future research and clinical practice will be presented (see Chapter 7).

Methods



Chapter 2

The Outcome Questionnaire (OQ-45) in a Dutch population: a cross-cultural validation

De Jong, K., Nugter, M.A., Polak, M.G., Wagenborg, J.E.A., Spinhoven, P. & Heiser, W.J. (2007).

Clinical Psychology and Psychotherapy, 14, 288–301.

Abstract

The cross-cultural validity of the Outcome Questionnaire (OQ-45) in the Dutch population has been examined by comparing the psychometric properties and equivalence in factor structure and normative scores of the Dutch OQ-45 with the original American version. Data were collected from a university (n = 268), in a community (n = 810) and from three mental health care organizations (n = 1920). Results show that the psychometric properties of the Dutch OQ-45 were adequate and similar to the original instrument. Some differences in equivalence were found though. In factor analysis, two additional factors were found: one consisting of social role items and another that reflected anxiety and somatic symptoms. Furthermore, normative scores were different for the Dutch and American samples, and this resulted in different cut-off scores for estimating a clinically significant change in the Dutch population.

Introduction

Over the last years, the Outcome Questionnaire (OQ-45; Lambert et al., 1996) has become one of the 10 instruments most frequently used by practitioners in the USA to measure clinical outcomes (Hatfield & Ogles, 2004) and is often used in clinical outcome research. It is also gaining popularity in other countries and has been translated into several languages, including Japanese, Korean, Italian, French, Portuguese, German and Dutch. Even though the psychometric properties of the original version of the OQ-45 have been thoroughly investigated, few articles are available on the properties of translated versions of the OQ-45. This article addresses the cross-cultural validity of the Dutch OQ-45.

Reasons for the popularity of the OQ-45 lie in the fact that it has some advantages that most other instruments do not have. First, the OQ-45 aims to measure three domains of functioning: symptom distress (SD), interpersonal relations (IR) and social role (SR) performance. It has become accepted in outcome research to measure symptom reduction as well as an improvement in well-being (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999). Other popular outcome instruments such as the Symptom Checklist-90 (SCL-90; Derogatis, 1977), the Brief Symptom Inventory (Derogatis, 1975) or the Social Adjustment Scale (SAS; Weissman & Bothwell, 1976) measure either symptoms or functioning. Additional instruments are required to measure the other domains. Moreover, the OQ-45, along with instruments such as the Clinical Outcomes in Routine Evaluation-Outcome Measure (Evans et al., 2000), is a general instrument that can be used for a large variety of disorders, so a comparison in functioning and outcome of a broad range of patients is possible, irrespective of psychiatric diagnosis. By contrast, specific instruments are designed for certain disorders and cannot compare symptoms and functioning of different types of patients.

Also, the OQ-45 is relatively short. An average patient can complete it in approximately 5 minutes. This is especially important in clinical practice, where there is neither time nor budget for the test batteries that are common in academic research. Probably the most important feature of the OQ-45 is that it is capable of tracking patient progress by repeated measurements. The OQ-45 is frequently used in outcome research that provides weekly feedback to the therapist about the patient's progress. The patient's treatment course is compared to a predicted course, and the therapist is alerted when the patient goes too far off the predicted track. This feedback results in more effective treatments, especially for those patients who have a higher risk for treatment failure (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005; Lambert et al., 2001, 2003).

Psychometric Properties

The psychometric properties of the American version of the OQ-45 have been extensively investigated. Reliability and validity estimates are good for the SD subscale and total scale. Reliability is adequate for the IR and SR subscales, but concurrent validity estimates are less convincing. The OQ-45 has proper sensitivity and specificity, and the sensitivity to change is good on the scale as well as on the item level (Lambert et al., 2004). Normative scores of the clinical and community samples differ significantly. The structure of the instrument, with three subscales, does not seem to have sufficient empirical support. Mueller, Lambert, and Burlingame (1998) found that the three-factor model did not have a proper fit in a confirmatory factor analysis. Chapman (2003) could not find support for the three-factor structure either and performed an exploratory analysis, which resulted in nine factors. However, these factors have not been confirmed in a new sample.

Equivalence between Language Versions

During translation, many problems may occur that change the properties of an instrument. First, there may be a difference in the meaning of the items of the instrument in a different language. Flaherty et al. (1988) refer to this as semantic equivalence and state that the key to obtain semantic equivalence is the back-translation method, which has been used in all OQ-45 translations. Still, a slight semantic difference is likely to occur, as the English language has considerably more words than the Dutch.

In order to assure the equivalence of two language versions of psychological tests, the constructs underlying the test need to be equivalent (Butcher, Derksen, Sloore, & Sirigatti, 2003). Cultural differences influence the conceptual equivalence of two language versions of an instrument. Hofstede (2006) introduced five cultural dimensions that can assist in differentiating between cultures. His research showed that American and Dutch cultures have similar levels of individualism, power distance and uncertainty avoidance, but differ in long-term orientation and masculinity.

One method of examining conceptual equivalence is to determine whether the items and scales maintain generally the same factors in the new language version. A well-known example of a difference in conceptual equivalence is the SCL-90: the Dutch version has a different factor structure than the original version (Arrindell & Ettema, 1975). Differences in normative scores often exist between different cultures, even among Western countries. For example, in the European Psychiatric Services: Inputs Linked to Outcomes and Needs (EPSILON) study, van Wijngaarden et al. (2000) found differences in scoring and reliability estimates on the Involvement Evaluation Questionnaire: scores were usually high in Verona and low in Copenhagen. The Minnesota

Multiphasic Personality Inventory (MMPI) also has different norms for different countries (Butcher et al., 2003). More importantly, a difference in normative scores may result in different criterion validity, also changing the sensitivity and specificity of the instrument. Flaherty et al. (1988) refer to this as a lack of criterion equivalence.

In clinical outcome research, a common criterion is whether clinical significant change (Jacobson & Truax, 1991) has occurred between post and pretreatment measurements. Because the cut-off point for estimating clinical significant change is based on population curves, it is important to take proper samples of these populations in the new culture. If criterion equivalence is not reached, a calibration of scores should be performed, thus calculating new cut-off scores for the population (Flaherty et al., 1988). Differences in cut-off scores are usually the result of significant differences in normative mean scores between two cultures, but may also occur when differences in mean scores are small (non-significant). Furthermore, differences in reliability estimates can lead to different reliable change indices (RCIs), because in the calculation of the RCI, the reliability of the instrument is used to estimate the measurement error.

The goal of the current research project was to find out if the factor structure and normative scores are equivalent with the original version and to determine the psychometric properties of the Dutch OQ-45. Preliminary studies showed that Dutch subjects seemed to have lower scores on the OQ-45 than American subjects (de Jong & Nugter, 2004). Also, some differences in psychometric properties were found. At that time, however, sample sizes were not large and representative enough to draw definite conclusions. Because the normative sample size was insufficient for individual use of the OQ-45 and more information was needed on the validity, more data were collected and additional studies were performed. For this paper, the data of initial and further studies of the OQ-45 are combined.

Method

Data

Subjects included a university sample, two community samples and two clinical samples. All data were collected for research purposes only. A student sample of 268 undergraduates was collected from the psychology department of the University of Amsterdam as part of its educational program. Subjects completed an OQ-45, SCL-90 and Groningse Vragenlijst Sociaal Gedrag-45 (GVSG-45) at the first session. A group of 264 students completed the OQ-45OQ-455 for the second time after 2 weeks. Data from the student sample were used only to calculate reliability and validity, as this group was not representative to function as a normative group.

A community sample of 446 individuals was collected by a random selection

of subjects in the phone directory of 13 phone regions, stratified by province, geographically distributed through the Netherlands and including major cities as well as suburban and rural regions. All adults in the household were asked to complete the questionnaire. If they consented to participate, questionnaires and consent forms were mailed. A selection of 1270 numbers was made of which 286 were never reached and 487 households consented. A total of 818 questionnaires were sent (with an average of 1.7 per household), and 446 were returned completed (55%).

Another community sample of 362 individuals was collected from 14 commercial and non-profit business settings in a variety of business branches. The questionnaires were distributed by internal mail. A stamped addressed envelope was included. Completion of the test was on a voluntary basis and was anonymous. A total of 1097 questionnaires was spread at the business settings, so the response rate was 33%, which is average for this type of sample. Subjects in the community sample who received treatment for psychological or psychiatric problems were deleted from the sample; 24 subjects were removed on this ground.

The first clinical outpatient sample of 1545 persons was collected at four sites of three mental health care institutions. Patients completed a paper-and-pencil version of the OQ-45 either before or after intake. To obtain test-retest stability data, a subsample of 43 patients who completed the OQ-45 after the intake completed the OQ-45 2-3 weeks later, before they received a treatment advice. During that time period, no treatment sessions took place. A different subsample of 117 patients completed the SCL-90 and Depression Anxiety and Stress Scale (DASS) together with the OQ-45 after intake to obtain concurrent validity estimates. By means of an internet screening tool, an additional sample of 375 patients completed the OQ-45 online prior

Table 1 Characteristics of the samples

Sample	*n*	Gender		Age	
		Female	Male		
		n (%)	*n (%)*	*Range*	*Mean (SD)*
Community sample	810	513 (63)	297 (37)	18-94	44.3 (15)
- Phone directory sample	446	247 (55)	199 (45)	18-94	48.1 (16)
- Business sample	361	264 (73)	97 (27)	18-77	39.5 (12)
University sample	268	171 (64)	96 (36)	17-53	22.3 (6)
Clinical sample					
- Outpatient sample paper-and-pencil	1545	896(58)	628(41)	18-65	37.3 (11)
* test-retest	42	31 (74)	11 (26)	18-55	31.7 (10)
*concurrent validity SCL-90/DASS	118	76 (64)	41 (35)	18-61	33.6 (11)
* sensitivity to change	60	32 (53)	24 (40)	23-62	41.6 (10)
- Outpatient internet screening tool	375	-	-	-	-

to their intake session. Depending on a global screening procedure, additional specific questionnaires that matched the subject's symptoms were completed by the subjects. An average number of 9.6 (*SD* = 3.3) questionnaires were completed by the patient.

An overview of sample characteristics is given in Table 1. There are some differences between the community and business samples with regard to gender and age, because one of the companies in the business sample, a private home care organization, had relatively many females among their employees. Data from this company were kept in the sample because scores on the OQ-45 did not differ significantly from the other business sites. For technical reasons, demographic characteristics of the internet screening sample were not available.

Subjects who left more than 20% of the questions of the questionnaire unanswered were removed from the sample; 8 students, 4 persons from the community sample and 49 outpatients were deleted. In case of missing values, a mean score for the remaining scale items was calculated, multiplied by the number of items on the scale and rounded to the nearest number. Mean subscale scores were only calculated if no more than 20% of the scale items were missing. Missing values were not replaced in those analyses that make use of the data on the item level (e.g., factor analysis, reliability analysis).

Instruments

The OQ-45

The OQ-45 consisted of 45 items that were scored on a five-point rating scale, ranging from never (0) to almost always (4). The SD subscale had 25 items that were associated with most common disorders in public mental health care; depression, anxiety and addiction to alcohol or drugs were well represented. The IR subscale consisted of 11 items and measured the functioning of the patient in relationships with partner, family and friends. The SR subscale contained nine items and measured functioning in school, work and leisure. There were nine reversely scored items.

The American normative sample consisted of undergraduate, community and clinical subsamples. The undergraduate samples were collected from universities in three states; the community sample was collected from various business locations and from the Utah phone directory, and the clinical samples were collected from a university counseling centre, an employee assistance program, a university-based outpatient clinic and a community mental health service centre (Lambert et al., 2004). Comparisons with the Dutch samples were made with the undergraduate, community and outpatient samples.

Instruments Used for Validation of the OQ-45

A short description of the instruments used for validation of the OQ-45 is given in this section. All instruments are self-rating questionnaires and have proper psychometric properties. References for both the original and the Dutch versions are given. A distinction is made between general and specific instruments.

General Instruments

The SCL-90-item version (Arrindell & Ettema, 1975) and DASS (de Beurs, van Dyck, Marquenie, Lange, & Blonk, 2001; Lovibond & Lovibond, 1995) were used to validate the SD subscale. For the SCL-90, the Global Severity Index (GSI) was calculated. For the DASS, subscale scores were used to correlate with the OQ-45 subscales.

The GVSG-45-item version (de Jong & van der Lubbe, 2001) was used to validate the IR and SR domain scores of the OQ-45. It measures social behaviour on nine domains. For this research, two indices that were not in the original questionnaire were calculated: as an index of interpersonal problems, we used the mean score of the Parents, Partner, both Children domains and the Friends domain, further referred to as Functioning on Interpersonal Relationships. As an index of SR performance, we calculated the mean score of the School, Occupation, Housework and Leisure domains, further referred to as Functioning on Social Role.

Specific Instruments

The OQ-45 claims to be applicable for a variety of disorders. Therefore, a number of instruments that aim to measure symptoms of specific disorders were compared with the OQ-45. The specific instruments were part of an internet screening tool. Not all instruments that were in the internet screening tool were used in analysis. The instruments were selected using two criteria. First, the number of patients who completed the instrument had to exceed 30. Second, the instruments had to measure symptoms that occur in a variety of patients, such as anxiety, depression, grief and reactions to overwhelming experiences. One would expect medium-high correlations between the specific instruments and the SD subscale of the OQ-45 and low-to-medium correlations with the IR and SR subscales.

The Quick Inventory of Depressive Symptoms Self-Report (QIDS-SR16; Rush et al., 2003) assesses all the criterion symptoms that the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition indicates to diagnose a major depressive episode.

The Body Sensations Questionnaire (BSQ) and Agoraphobic Cognitions Questionnaire (ACQ) (Bouman, 1995; Bouman, 1998; Chambless, Caputo, Bright, & Gallager, 1984) both measure experienced anxiety during panic or anxiety attacks. The BSQ measures anxiety for physical sensations, while the ACQ measures catastrophic cognitions during the anxiety or panic attack.

The Liebowitz Social Anxiety Scale-Self-Report (LSAS-SR; Liebowitz, 1987; van Balkom, de Beurs, Hovens, & van Vliet, 2004) measures the severity of social phobic symptoms and avoidance behaviour.

The Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990; van Rijsoort, Vervaeke, & Emmelkamp, 1999) measures the inclination to worry as well as the amount, intensity and uncontrollability of worrying.

The Padua Inventory-Revised (Sanavio, 1988; Van Oppen, Hoekstra, & Emmelkamp, 1995) measures obsessive thoughts and compulsions on five domains: impulse, wash, control, ruminate and precision. The total score is used to assess the severity of obsessive-compulsive symptoms.

The Impact of Events Scale Revised (IESR; Weiss & Marmar, 1997) consists of 22 items measuring reexperience of shocking events and avoidance of thoughts and feelings that are related to these events. It also measures increased arousal. There are three subscales: Intrusion, Avoidance and Hyperarousal. The Hyperarousal subscale was not yet normed for the Dutch population and is therefore not included in the total score. The Dutch translation of the IESR is known as the Schokverwerkingslijst (Brom & Kleber, 1985; van der Ploeg, Mooren, Kleber, van der Velden, & Brom, 2004).

The Inventory of Complicated Grief-revised (ICG-r; Prigerson, Kasl, & Jacobs, 1997) measures symptoms of normal and potentially complicated grief and mourning. The Dutch translation of the ICG-r is known as the Rouw Vragenlijst (Boelen, de Keijser, & van den Bout, 2001).

Analysis

The paper-and-pencil data were scanned by computer and validated using the Teleform software from Verity, Inc.; Sunnydale, CA; USA, version 9.0. Data from the internet screening tool were collected with the screening tool developed specifically for one of the mental health care organizations by Interapy (2004).

All tests of difference were two-tailed against $p < 0.05$. The large sample sizes gave high statistical power; therefore, effect sizes were also reported. Effect sizes (Cohen's *d*) were calculated with the effect size spreadsheet by Thalheimer and Cook (2002) in Microsoft Excel and were interpreted according to the criteria reported by Cohen (1992). Most analyses were conducted in SPSS for Windows, version 13.0 (2004) SPSS Inc. Chicago, IL: USA. The confirmatory factor analysis was performed in EQS 6.1 (2005), Multivariate Software Inc., Encino, CA: USA. Exploratory principal component analysis on the residual matrix was performed in MATLAB, release 14 (2005) The Mathworks Inc., Natick, MA: USA.

The goodness-of-fit indices that were reported in the confirmatory factor analysis were the root mean square residual (RMR), the root mean square error of approximation

(RMSEA), the GFI, the Bentler-Bonnet normed fit index (NFI), the comparative fit index (CFI), the chi-square (χ^2) and the chi-square divided by degrees of freedom in the model (χ^2/df). General guidelines were that the RMR should be less than 0.10, the RMSEA less than 0.05, the GFI greater than 0.95, the CFI and NFI greater than 0.90, a χ^2 that is non-significant and χ^2/df less than 2 (Kline, 1998). The goodness-of-fit indices that were reported in the original study by Mueller et al. (1998) are also reported in the current paper, except for the adjusted goodness of fit index and critical *n*. The RMSEA was added.

RESULTS

Equivalence

Conceptual Equivalence

Factor Analysis

As was stated earlier, the factor structure of an instrument may change in a different cultural setting or language. The sample, consisting of the community and clinical sample, was randomly split in two so that if additional analyses would result in new factor structures, they could be fitted on the other split of the sample to test the stability of the factor solution. The first sample was used to test whether the OQ-45 had a three-factor structure. A confirmatory factor analysis was conducted, using general least squares estimation to make our solution comparable to the original analyses of Mueller et al. (1998).

As can be seen in Table 2, our solution meets three out of seven goodness-of-fit criteria. The indices that did not meet the criteria were the χ^2, χ^2/df, RMR and GFI. The χ^2 and χ^2/df are dependent on the sample size, and given our large sample size and consequently high power, a significant χ^2 is not necessarily a sign of a poor fit. The criteria for a good fit for the RMSEA, NFI and CFI criteria are met and the three-factor solution seems to have a reasonable fit. The fit of our solution is notably better than the solution that Mueller et al. (1998) obtained; their three-factor solution met none of the criteria that we applied.

Table 2 Confirmatory factor analysis Goodness of Fit Indices

Model	χ^2	*df*	χ^2/df	RMR	GFI	AGFI	NFI	CFI
Three-factor solution	3678.2	942	3.90	.103	.880	.868*	.933*	.949*
Five-factor solution	3413.4	925	3.69	.075*	.889	.875*	.957*	.964*

*RMR = Root Mean square Residual; GFI = Goodness-of-Fit Index; AGFI = Adjusted Goodness-of-fit Index; NFI = Normed Fit Index; CFI = Comparative Fit Index; * Meets the recommended criteria*

Table 3 Standardized factor loadings of the factor models

	3-factor solution (n = 1362)			*5-factor solution (n = 1363)*				
Item	F1	F2	F3	F1	F2	F3	F4	F5
2	.66			.65				.16
3	.63			.62				
5	.71			.73				
6	.73			.72				
8	.67			.66				
9	.85			.82				.13
10	.78			.67				.38
11	.18			.14				
13	.84			.87				
15	.84			.87				
22	.68			.68				
23	.80			.83				
24	.75			.77				
25	.76			.75				.14
27	.45			.36				.34
29	.61			.51				.44
31	.86			.88				
33	.71			.61				.32
34	.37			.33				.39
35	.48			.38				.39
36	.73			.68				.29
40	.73			.71				
41	.57			.56				.25
42	.84			.87				
45	.45			.41				.34
1		.59			.60			
7		.60			.60			
16		.45			.41			.21
17		.46			.44			
18		.81			.81			
19		.58			.62			
20		.71			.76			
26		.27			.27			
30		.74			.73			
37		.70			.71			
43		.80			.84			
4			.71			.51	.39	
12			.63			.68		
14			.13			-.11	.34	
21			.76			.73		
28			.34			.30		
32			.21			.20		
38			.81			.66	.50	
39			.65			.44	.48	
44			.66			.58		

Table 3 shows the standardized factor loadings on the three factors. Four items have factor loadings below 0.30: items 11, 14, 26 and 32. Items 11, 26 and 32 are known difficult items. They all measure problematic alcohol/drug use and have rather skewed scoring (a lot of '0' scores). In exploratory factor analysis, they consequently end up in one factor, which consists only of these three items. Item 14 'I work/study too much' is another special case. This item does not perform well in the original OQ-45 also and shows a negative correlation with several items in the covariance matrix. Moreover, it is the only item in the SR subscale wherein the functional sample ($M = 1.87$, $SD = 1.1$) actually scores slightly higher than the clinical sample ($M = 1.68$, $SD = 1.2$), $t(1706) = 3.54$, $p < 0.001$, $d = 0.14$.

As the three-factor solution still was not satisfactory, we tried to explain more variance by using the residual matrix from the three-factor solution. A principal component analysis with varimax rotation produced two components with an eigenvalue greater than 1, which explained 34% of variance. Because of the varimax rotation, the two components are not correlated, and as the solution was based on the residuals of the three-factor solution, they are considered independent from the first three factors as well. For each component, the items that loaded above 0.15 were selected.

The two components were added to the original three-factor model and fitted on the second split sample. The chi-square value of the five-factor solution has dropped 264.7 points, associated with a loss of 17 degrees of freedom, which leads to a χ^2/df improvement ratio of 15.6. A ratio of 2 is usually considered a substantial improvement. The other goodness-of-fit values also show an improvement of fit. Especially important is the RMR value, which did not meet the criterion of 0.10 in the three-factor solution, but does so in the five-factor solution.

The first additional factor consists of items that all belong to the SR domain. The unexplained variance in the SR domain in the three-factor solution seems to be mainly caused by item 14, which has a low factor loading on this domain. In the five-factor solution, this item even has a negative factor loading on the SR domain. For clinical purposes, the extra social factor does not add much. It may be useful if one is merely interested in the social role functioning of an individual, but in that case, the OQ-45 would probably not be used as the instrument of choice.

In contrast, the second additional factor does seem to add something to the clinical utility of the instrument. Most of the items on this factor originate from the SD scale, which is a rather long scale. The items seem to be related to anxiety and somatic manifestations of anxiety. Some of the items represent cognitive representations of anxiety, such as item 10, 'I feel fearful', whereas others seem more physical manifestations of anxiety, such as item 29, 'My heart pounds too much', which is known to be a symptom of anxiety or panic attack. It may be a useful addition to the original

three-factor structure. Therefore, we decided to evaluate the validity of this factor, along with the validity of the original three factors. The factor is further referred to as Anxiety and Somatic Distress (ASD).

Correlations between Subscales

The correlations between the subscales of an instrument give an indication of whether the structure of the instrument is as it was intended. In the case of an instrument that assesses several domains of functioning, multidimensionality should be reflected in the factor structure of the instrument. Also, each subscale should assess a concept that is not measured by the other subscales. Therefore, the correlations between the domains should not be too high. Table 4 shows that the correlation between the subscales of the OQ-45 is higher than is desirable, indicating a moderate construct validity. Especially high is the correlation between the SD and ASD subscales. This is not surprising, as the ASD subscale consists almost exclusively of items that are in the SD scale, but considering that, correlations would ideally be lower.

Criterion Equivalence

Differences in Scoring

Table 5 shows the mean scores of Dutch and American samples on the OQ-45 subscales and total scale. The mean scores for the Dutch community (t (1620) = 7.48, $p < 0.001$, $d = 0.37$) and clinical samples (t (2260) = 2.50, $p = 0.01$, $d = 0.15$) are somewhat below the American equivalents, even though the effect sizes are small. The mean scores of Dutch students are somewhat higher than the American student sample (t = -4.40, $p < 0.001$, $d = 0.39$). Figure 1 gives a visual representation of the sample differences for the OQ-45 total scale.

In the American samples, no differences were found between males and females. In the Dutch samples, some small differences were found. In Table 6, scores for the clinical and community samples are given for males and females. In the community

Table 4 Correlations between the subscales and total scale

	SD	ASD	IR	SR	Total
Symptom Distress (SD)	1 ($n = 2726$)				
Anxiety and Somatic Distress(ASD)	.94 ($n = 2726$)	1 ($n = 2726$)			
Interpersonal Relations (IR)	.75 ($n = 2723$)	.64 ($n = 2723$)	1		
Social Role (SR)	.68 ($n = 2646$)	.60 ($n = 2646$)	.59 ($n = 2644$)	1	
OQ-45 Total score	.97 ($n = 2726$)	.89 ($n = 2726$)	.85 ($n = 2724$)	.77 ($n = 2647$)	1 ($n = 2727$)

sample, significant differences were found for gender, Wilks' $\lambda = 0.92$, $F(5, 792) = 13.3$, $p < 0.001$. Women showed higher levels of SD, ($F[1, 796] = 10.7$, $p = 0.001$, $d = 0.26$) and ASD ($F[1, 796] = 21.8$, $p < 0.001$, $d = 0.37$), while men showed more problems in SR performance ($F[1, 796] = 13.6$, $p < 0.001$, $d = 0.27$).

Similar results were found in the clinical sample: Wilks' $\lambda = 0.92$, $F(5, 1446) = 23.7$, $p < 0.001$. Here, women showed slightly higher levels of SD ($F[1, 1450] = 5.83$, $p = 0.016$, $d = 0.13$) and ASD ($F[1, 1450) = 29.1$, $p < 0.001$, $d = 0.29$), while men showed somewhat more problems in SR performance ($F[1, 796] = 25.4$, $p < 0.001$, $d = 0.27$).

Clinical Significance and Reliable Change

To measure individual change, the criterion of clinical significance by Jacobson and colleagues is often applied (Jacobson & Truax, 1991; Jacobson, Follette, & Revenstorf, 1984; Jacobson, Roberts, Berns, & McGlinchey, 1999). Their criterion is twofold: (1) the magnitude of change has to be statistically reliable, and (2) by the end of treatment, patients have to end up in a (score) range that renders them indistinguishable from well functioning people. A cut-off point for clinically significant change and an RCI are calculated using formula c by Jacobson and Truax (1991).

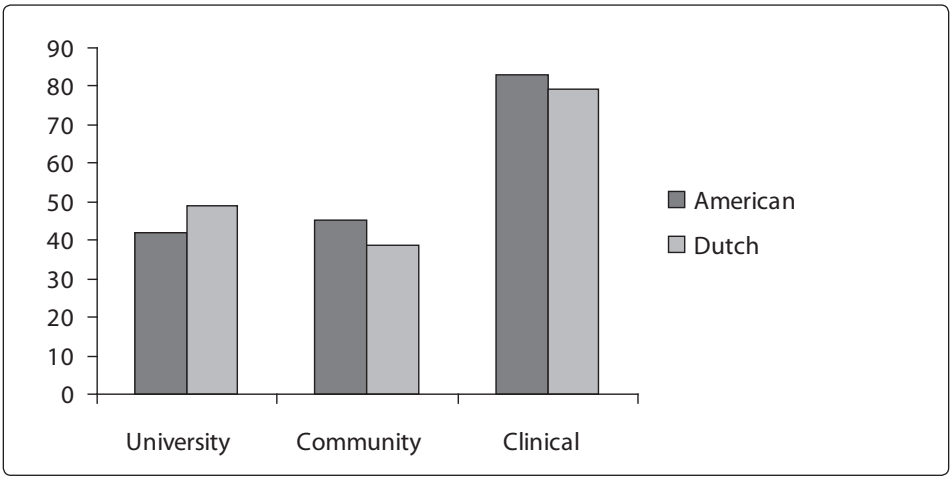
For the American OQ-45, the cut-off score for clinical dysfunctioning is 63 for the total scale and 36, 15 and 12, respectively, for the SD, IR and SR subscales. The RCI is 14 for the total scale and 10, 8 and 7 for the subscales, respectively (Lambert et al., 2004). For the Dutch OQ-45, the cut-off score for the SD subscale is 33; for the ASD subscale, it is 19; for the IR subscale, it is 12; for the SR subscale, it is 10, and for the total scale, the cut-off score is 55. A person that scores on or above the cut-off score belongs to the dysfunctional (clinical) range. Given the differences found between male and female respondents, separate cut-off scores for men and women were also calculated for the

Table 5 Means and standard deviations of OQ-45 in the Dutch and American samples

	American samples						Dutch samples					
	University ($n = 235$)		Community ($n = 815$)		Clinical ($n = 342$)		University ($n = 268$)		Community ($n = 807$)		Clinical ($n = 1920$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Symptom Distress	23.0	10	25.4	12	49.4	15	27.3	12	22.2	10	48.9[b]	16
Anxiety and Somatic Distress	-	-	-	-	-	-	15.6	7	13.3	6	25.9[b]	9
Interpersonal Relations	8.8	5	10.2	6	19.7	6	11.4	5	8.4	5	16.8[c]	7
Social Role	10.4	4	9.6	4	14.1	5	10.4	4	8.1[a]	3	13.6[d]	6
Total score	42.2	17	45.2	19	83.1	22	49.1	18	38.7	16	79.5	25

Notes: Means and standard deviations of the American samples were copied from the OQ-45 manual (Lambert et al, 2004);
[a] 798 cases; [b] 1919 cases; [c] 1917 cases; [d] 1849 cases

Figure 1 OQ-45 total score for the American and Dutch samples



subscales with gender differences. The cut-off score for the SD subscale was 31 for men and 33 for women; the cutoff for the ASD subscale was 17 for men and 19 for women, and the cut-off for the SR subscale was 12 for men and 10 for women.

Using the cut-off score of 55, sensitivity for the OQ-45 total scale is 0.84, which means that 84% of the community sample is correctly identified as belonging to the functional sample. The specificity of the OQ-45 is 0.85, which means that 85% of the clinical sample is correctly identified as dysfunctional. Using the cut-off score of 63 from the original OQ-45 leads to a higher sensitivity (0.93), but lower specificity (0.74). This means that fewer patients are correctly identified as belonging to the dysfunctional sample.

The RCI is usually expressed as the amount of points on a certain measurement instrument that a patient has to improve between preand posttreatment measurements. The RCI depends on the reliability of the measurement instrument and the variability of scores. As reliability index, the pooled internal consistency of the

Table 6 Means and standard deviations by gender in the clinical and non-clinical sample

	Community sample						Clinical sample					
	Male			Female			Male			Female		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Symptom Distress	296	20.6	10	511	23.2	10	628	47.4	15	896	49.3	16
Anxiety and Somatic Distress	296	11.9	6	511	14.1	6	628	24.3	9	896	26.8	9
Interpersonal Relations	296	8.3	5	511	8.4	5	627	16.6	7	894	16.6	7
Social Role	292	8.6	4	506	7.7	3	598	14.3	5	857	12.8	5
OQ-45 Total score	296	37.4	16	511	39.4	16	628	78.4	25	896	79.0	25

clinical and community sample was used (see Table 7). Using either subgroup would lead to less variability in the answers, which leads to lower values of Cronbach's alpha. In literature, this phenomenon is referred to as range restriction (Cronbach, 1990). The RCIs for the SD, ASD, IR and SR subscales are 10, 8, 8 and 9, respectively. The RCI for the OQ-45 total scale is 14, so a patient has to improve a minimum of 14 points on the OQ-45 to obtain reliable change.

Psychometric Properties

Reliability

Internal consistency estimates are sufficient for subscales and the total scale in most of the samples (see Table 7), except for the SR subscale, for which disappointing values for Cronbach's alpha were found in the university, community and clinical samples. Combining the clinical and community samples improves the results, which indicates that restriction of range may occur here. Another explanation may be that an increased sample size improves internal consistency values. In reliability analysis, cases are rapidly lost: if one item of the scale is missing, the case cannot be used entirely.

We tried replacing missing values with the mean score of the remaining scale items. This resulted in slightly better a values for the clinical sample, but not for the community and the university samples. Beside restriction of range, the SR subscale has the lowest number of items, and two of the items that were awkward in the three-factor factor analysis belong to this scale. These two items, items 14 and 32, have relatively low item-total correlations (r_{it} = 0.11-0.15).

Most values are similar to values that were found in the American sample. No reports of internal consistency in the American community sample exist, so comparison is not

Table 7 Internal consistency (Cronbach's α) and test-retest reliability (Pearson's product-moment correlation coefficient)

	Internal consistency								Test-retest	
	University		Community		Clinical		Community and clinical		University (n = 264)	Clinical (n = 42)
Domain	*n*	*α*	*n*	*α*	*n*	*α*	*n*	*α*	*r*	*r*
Symptom Distress	257	0.90	768	0.89	1247	.91	2390	0.95	0.81	0.76
Anxiety and Somatic Distress	261	0.79	786	0.82	1743	0.84	2529	0.89	0.74	0.70
Interpersonal Relations	264	0.74	770	0.77	1607	0.80	2377	0.84	0.71[a]	0.83
Social Role	258	0.61	773	0.53	1620	0.69	2393	0.72	0.73	0.74
OQ-45 Total score	247	0.92	726	0.91	1309	0.93	2035	0.96	0.82	0.79

Note: [a] 262 cases

Table 8 Concurrent validity estimates for the OQ-45 with SCL-90, DASS and GVSG-45

	Clinical ($n = 118$)				University ($n = 268$)		
	GSI	DASS-D	DASS-A	DASS-S[a]	GSI	FIR	FSR
Symptom distress	**0.80**	**0.78**	**0.74**	**0.72**	**0.78**	0.42	0.54
Anxiety and Somatic Distress	**0.75**	0.63	**0.74**	0.60	**0.66**	0.34	0.42
Interpersonal relations	0.62	0.54	0.38	0.54	0.59	**0.51**	0.51
Social role	0.51	0.51	0.46	0.48	0.57	0.38	**0.55**
OQ-45 Total score	0.80	0.77	0.68	0.72	0.77	0.49	0.60

Note: GSI = Global Severity Index of the Symptom Checklist 90 – Revised (SCL-90-R); DASS-D = Depression Anxiety Stress – Depression subscale; DASS-A = Depression Anxiety Stress – Anxiety subscale; DASS-S = Depression Anxiety Stress – Stress subscale; FIR = Functioning on interpersonal relationship, based on the Groningse Vragenlijst Sociaal Gedrag 45-item version (GVSG-45) subscales Parents, Partner, Children and Friends; FSR= Functioning on social role, based on the GVSG-45 (Groningen Questionnaire of Social Behaviour) subscales Study, Work, Housework and Leisure. All correlations are significant at the .01 level. [a] 117 cases

possible, but in a German community sample, Cronbach's alpha for the SR subscale has been found to be 0.59, which is close to our value of 0.53 (Lambert, Hannöver, Nisslmüller, Richard, & Kordy, 2002).

Test-retest reliability is an indication for the stability of scoring over time. Very marked score changes over a short period of time would be problematic. The correlation between the first and second completion of the OQ-45 is sufficient for both clinical ($r_{tt} = 0.70$-0.83) and student ($r_{tt} = 0.71$-0.81) samples.

Validity

Criterion Validity. An important validity requirement of an outcome measure is that it should discriminate between the clinical population for which it is designed and the functional (community) population. Table 8 shows that the difference between community and clinical means were large.

The community sample has a highly significantly better level of functioning on all subscales and the total scale, Wilks' $\lambda = 0.58$, $F(5, 2637) = 388.1$, $p < 0.001$. Effect sizes for the difference between the clinical and community samples are very large for the IR ($F[1, 2643] = 960.5$, $p < 0.001$, $d = 1.32$) and SR ($F[1, 2643] = 674.1$, $p < 0.001$, $d = 1.10$) subscales, and huge for the SD ($F[1, 2643] = 1873.2$, $p < 0.001$, $d = 1.83$) and ASD ($F[1, 2643] = 1280.7$, $p < 0.001$, $d = 1.52$) subscales and total scale ($F[1, 2643] = 1804.5$, $p < 0.001$, $d = 1.80$).

Concurrent Validity

To assess the concurrent validity, three subsamples completed additional questionnaires together with the OQ-45. Results are presented in Table 8. The SCL-90 and DASS were used to validate the SD and ASD subscales. The concurrent validity of the SD subscale with the GSI of the SCL-90 was slightly below the American value in the

clinical sample ($r = 0.80$ versus $r = 0.84$), but better in the university sample (0.78 versus 0.61). The correlations between the SD and DASS subscales were adequate: neither too high, nor too low. The ASD subscale also showed proper concurrent validity with the SCL-90 and the Anxiety subscale of the DASS ($r = 0.74$). Correlations between the Depression ($r = 0.63$) and Stress ($r = 0.60$) subscales and the ASD subscale were lower, as was to be expected.

It was difficult to find a Dutch instrument to validate the IR and SR subscales, and we had to calculate our own indices with the instrument we finally used. Nonetheless, the convergent validity of the GVSG-45 with the IR ($r = 0.51$) and SR subscale ($r = 0.55$) falls in the range of correlations of the OQ-45 subscales with the Inventory of Interpersonal Problems ($r = 0.49$-0.64) and Social Adjustment Scale ($r = 0.44$-0.73), respectively in the American samples.

Correlations with other Psychological Constructs. In the internet screening tool sample, several instruments that measure specific disorders were administered together with the OQ-45 (see Table 9). On almost all specific questionnaires validity estimates are good, showing high correlations with the SD subscale (and subsequently, the OQ-45 total scale) and lower correlations with the IR and SR subscales. Exceptions are the PSWQ and the Quick Inventory of Depressive Symptoms-Self-Report 16 (QIDS-SR16). The correlation between the PSWQ and the SD subscale is not high ($r = 0.38$), even though it is lower for the ASD, IR and SR subscales, as expected. The concept of worrying may not be uniquely linked to a certain pattern in symptoms.

The QIDS-SR16 shows a good concurrent validity with the SD subscale ($r = 0.78$), but correlations with the ASD ($r = 0.65$), IR ($r = 0.47$) and SR subscales ($r = 0.44$) seem higher than desirable. The ASD subscale shows good concurrent validity on the ACQ

Table 9 Correlations for the OQ-45 with instruments measuring other psychological constructs

	N	Symptom Distress	Anxiety and Somatic Distress	Interpersonal Relations	Social Role	OQ-45 Total score
ACQ	119	**0.58**	**0.62**	0.27	0.38	0.56
BSQ	119	**0.50**	**0.57**	0.22	0.24	0.46
ICG-r	56	**0.60**	0.62	0.33	0.38	0.60
IESR	110	**0.44**	0.47	0.11*	0.13*	0.35
LSAS-SR	54	**0.62**	**0.50**	0.48	0.34	0.63
PI-R	137	**0.57**	0.52	0.31	0.35	0.55
PSWQ	122	**0.38**	0.26	0.15*	0.18*	0.33
QIDS-SR16	164	**0.78**	0.65	0.47	0.44	0.77

Note: ACQ = Agoraphobic Cognitions Questionnaire; BSQ = Body Sensations Questionnaire; GL= Gewaarwordingenlijst; ICG-r = Inventory of Complicated Grief (ICG-r); IESR = Impact of Events Scale; LSAS-SR= Liebowitz Social Anxiety Scale- Self Report; PI-R = Padua Inventory-Revised; PSWQ = Penn State Worry Questionnaire; QIDS-SR16 = Inventory of Depressive Symptoms- Self Report 16 items version
* Non-significant correlations

(r = 0.62) and BSQ (r = 0.57) and the LSASSR (r = 0.50). Unexpected is the relatively high correlation of the ASD subscale with the ICG-r (r = 0.62). Of further interest is the correlation between the IR subscale and the LSAS-SR (r = 0.48). Having symptoms of social anxiety will probably influence interpersonal functioning, which shows a subsequently somewhat higher correlation.

In summary, the concurrent validity of the SD, ASD subscales and the total scale seems good, but validity estimates of the IR and especially SR subscales are less convincing (Table 9).

Sensitivity to Change

Another important criterion for instruments that are used for outcome and progress research is that they are capable of measuring changes in functioning that occur as a result of treatment. A subsample of 60 patients received a short treatment, with a maximum of five sessions. The OQ-45 was administered before and after treatment. The OQ-45 showed high sensitivity to change on all subscales (SD: $t(55) = 6.8$, $p < 0.001$, $d = 1.29$; ASD: $t(56) = 7.7$, $p < 0.001$, $d = 1.43$; IR: $t(51) = 4.3$, $p < 0.001$, $d = 0.84$; SR: $t(55) = 4.1$, $p < 0.001$, $d = 0.77$) and the total scale, $t(56) = 7.1$, $p < 0.001$, $d = 1.33$.

Discussion

This study investigated the cross-cultural validity of the OQ-45 in the Dutch population by evaluating the psychometric properties of the Dutch version and its equivalence with the original version of the OQ-45. The results show that the language versions are similar when it comes to reliability and validity estimates, but differences in factor structure and normative scores have been found.

The three-domain structure of the instrument, for which there was no strong evidence in the original version, had a reasonable fit in the Dutch population. Further analyses resulted in two additional factors that overlap mainly with the SR and SD subscales. The first one, which consisted of four items that are in the SR domain, was unexpected but not unexplainable considering the bad performance of item 14 ('I work/study too much'). This item has low item-total correlation and also came to notice in the reliability analysis. It is problematic in the original OQ-45 as well (see Mueller et al., 1998) and probably does not represent problematic behaviour. In fact, in contemporary society, some people may consider working too hard a good quality. The second factor, named ASD, was considered a useful addition to the existing scales and was therefore used in further analyses. Reliability and validity estimates for the ASD factor are promising. This factor may be especially interesting for use by care providers that specialize in anxiety or psychosomatic disorders.

Finding additional factors to the original structure does not seem to indicate

conceptual equivalence between the two versions of the OQ-45. And the fit of our solution is notably better than the fit of the original three-factor solution. However, this does not necessarily imply that they are not equivalent. Some of the GFIs that we found for the three-factor solution were similar to the ones reported in the American sample (Mueller et al., 1998). Running the same statistical analyses as we did may result in a similar structure in the American OQ-45.

The correlations between the subscales were too high. This suggests inadequate conceptual equivalence. Another possible explanation for the high correlations between the subscales may be that there really is a mutual interdependence between the concepts and that distress in one area influences functioning in other areas.

Comparison of normative scores between the American and Dutch populations showed that the Dutch community and clinical samples scored somewhat below their American equivalents. As was mentioned earlier, differences in scoring between culturally different populations are common in psychological testing. For instance, differences were found in the EPSILON study and on instruments such as the MMPI. The Dutch students showed higher scores than the American students. This may be due to a difference in sampling. In the Netherlands, the sample consisted of psychology students, whereas the American sample included other disciplines as well. Even though the differences between the populations were relatively small, calibration of cut-off scores and RCIs was necessary, for a lack of criterion equivalence occurred. Calibration resulted in a cut-off score for the Dutch population of eight points below the American cut-off point. After calibration, sensitivity and specificity values were very similar to those of the original version. The specificity of the American and Dutch OQ-45 is 0.83 and 0.85, respectively, and the sensitivity is 0.84 for both versions. The RCIs were equal as well.

A marked difference between the American and Dutch normative scores is that in the Dutch population, gender differences were found in both the clinical and the community sample. Men had more problems in the SR domain, whereas women showed higher levels of SD as well as ASD. Gender differences in normative scores are quite common in testing, and it is surprising that they were not found in the original OQ-45. The OQ-45 manual reports that in one study, some statistically different mean scores were found in the patient sample, but they were not considered to be of clinical relevance. In the OQ-45 manual, gender scores are only reported on a relatively small subsample, so a lack of power may be causing the insignificant findings.

When developing normative scores for any test, the quality of the population samples is very important. By combining phone book and business setting sampling, we strived for a representative sample of the Dutch functional population. This was more complicated for the clinical sample, as we could only address patients who received treatment in the participating mental health care centres. Given that the

sample size is large and multiple mental health care organizations participated, we believe our sample to be representative for the Dutch outpatient population.

Besides examining the equivalence, the psychometric properties of the Dutch version were investigated. The reliability of the subscales and the total scale was adequate in most of the samples. An exception was the internal consistency of the SR domain, which was too low in all three samples, but was substantially better when the clinical and community samples were combined. Sensitivity to change is very good, and the OQ-45 can effectively discriminate between functional and dysfunctional populations. The concurrent validity showed proper values for the SD and ASD subscales, but less support for the IR and SR subscales.

This study did not address the OQ-45 as a measure for tracking patient progress. More research should be conducted with the Dutch OQ-45 on this subject to obtain a better comparison of progress curves between the Dutch and American population. Given the differences in normative scores that were found in the current study, differences in treatment progress are to be expected. In connection with that, the sensitivity to change on a session to session basis should be investigated. Also, further research should be conducted on the SR domain and the additional factor that was identified in the present study. We are currently performing a pilot study with a different formulation of item 14 ('My work/study is too much for me') that may better reflect problematic functioning in the SR domain. The Dutch OQ-45 has moderate to good psychometric properties and is ready for use in clinical practice. However, separate norms for patient progress should be developed for the Dutch population.

Summarized, the Dutch OQ-45 has similar psychometric properties as the original instrument, but the two versions are not equivalent on all aspects. There may be a difference in conceptual equivalence, although further analyses with the original instrument can prove otherwise. Criterion validity of the Dutch OQ-45 was similar to the original values, but only after calibration of cut-off scores, indicating a lack of criterion equivalence. These results imply that a similarity in psychometric properties does not guarantee equivalence. The fact that calibration of cut-off scores was necessary, even though differences in population scores were small, shows the importance of proper normative scores for translated instruments. This is especially true for clinical outcome measures such as the OQ-45, where use of the 'wrong' norms may lead to faulty treatment decisions.

Chapter 3

A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects

De Jong, K., Moerbeek, M. & Van der Leeden (2010).

Psychotherapy Research, 20(3), 273-284.

Abstract

Multilevel analysis (or hierarchical linear modeling) has become increasingly popular for the analysis of longitudinal data in psychotherapy research. Over the last few years, three-level longitudinal models have become more common in psychotherapy research, particularly in therapist-effect or group-effect studies. Thus far, limited attention has been paid to power analysis in these models. This article demonstrates the effects of intraclass correlation, level of randomization, sample size, covariates and drop-out on power, using data from a routine outcome monitoring study. Results indicate that randomization at the patient level is the most efficient, and that increasing the number of measurements does not increase power much. Adding a covariate or having a 25% drop-out rate had limited effects on study power in our data. In addition, the results demonstrate that sufficient power can be reached with small sample sizes, but that larger sample sizes are needed to prevent estimation bias for the model parameters and standard errors.

Introduction

Multilevel analysis (Goldstein, 2003; Hox, 2002; Raudenbusch & Bryk, 2002; Snijders & Bosker, 1999) has gained significant popularity as a statistical technique for the analysis of longitudinal data in psychotherapy research over the last decade. It has been used to analyze growth models of phenomenon such as patient progress and expected treatment response (e.g. Finch, Lambert, & Schaalje, 2001; Haas, Hill, Lambert, & Morrell, 2002; Lueger et al., 2001; Lutz, 2002; Lutz, Martinovich, & Howard, 1999; Lutz, Rafaeli, Howard, & Martinovich, 2002; Slee, Garnefski, Van der Leeden, Arensman, & Spinhoven, 2008), the dose-response relationship (Lutz, Lowry, Kopta, Einstein, & Howard, 2001) and group therapy (e.g. Haringsma, Engels, Van der Leeden, & Spinhoven, 2006; Tasca, Balfour, Ritchie, & Bissada, 2007). Multilevel analysis describes the class of methods employing hierarchical regression models, and is also referred to as hierarchical linear modeling, linear mixed modeling, random effects regression modeling and random coefficient modeling. These models explicitly take into account the hierarchical structure of the data (the fact that repeated measurements are nested within patients). The popularity of multilevel analysis in analyzing longitudinal psychotherapy data lies in its flexibility to handle missing data and unbalanced designs, its capacity to model individual growth trajectories (Van der Leeden, 1998), the within-subject covariance structure (Hedeker & Gibbons, 2006), as well as models with three or more levels.

In higher level models, an additional level at which the subjects are clustered (e.g. 'therapist' or 'organization'), is added to account for between-therapist or between-site variance. Three level models are becoming more common in psychotherapy research, especially in studies on group therapy and therapist effects (e.g. Elkin, Falconnier, Martinovich, & Mahoney, 2006; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Nielsen, & Ogles, 2003).

For a special section on therapist effects in *Psychotherapy Research,* data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (TRDCP, Elkin et al., 1989) were re-analyzed using two multilevel approaches. Employing a two-level (non-longitudinal) model with functioning at termination as the dependent variable, Kim, Wampold & Bolt (2006) showed significant differences between therapists in treatment results. In contrast, a three-level longitudinal model by Elkin, Falconnier, Martinovich & Mahoney (2006) did not find significant therapist variance. These contrasting results initiated an exploration of whether three-level longitudinal models are appropriate for psychotherapy research. Wampold & Bolt (2006) posit that, in psychotherapy research, the level of functioning at treatment termination is important; not the change pattern, arguing a move away from longitudinal models. However, there are serious drawbacks to both the completers sample analysis they suggest and the missing data imputation methods used in intent-to-treat analysis that

longitudinal analysis does not have (Crits-Christoph & Gallop, 2006).

One of the problems in the re-analyses of the TRDCP data was that the sample size was smaller than is generally recommended for multilevel analysis (Kim, Wampold, & Bolt, 2006; Soldz, 2006). Given that therapist-effects have been found in analyses with larger datasets (Lutz, Leon, Martinovich, Lyons and Stiles, 2007) it is possible that power might have been an issue in the TRDCP analyses. Therapist effects have usually been small in randomized controlled trials and small to medium in naturalistic studies (Crits-Christoph & Gallop, 2006). It seems that therapist effects are harder to detect in randomized controlled trials using treatment manuals (Crits-Christoph et al., 1991) and that therapist variance declines when therapists are trained (David M. Clark, personal correspondence). Yet, even a small amount of variance at the therapist level, can have a significant influence. It has been shown that ignoring a level of nesting in the data can have considerable effects on the estimated variances and power to detect treatment or covariate effects (Moerbeek, 2004) and can seriously inflate the Type I error rate and size of the treatment effect (Wampold & Serlin, 2000). Numerous researchers have shown that ignoring clustering can also lead to serious errors in interpreting the results of statistical significance tests (e.g. Nich & Carroll, 1997). As a result, in many cases, adding a third level that models between-therapist variance is necessary, regardless of whether therapist effects are the main topic of interest in the study.

Having nested data can strongly influence the power to detect a treatment effect. Unfortunately, many researchers still use power analyses that ignore the effect of nesting, even if they plan to do a multilevel analysis. The reason for this is that in multilevel analysis (a priori) estimation of power is complex, as it depends on a number of variables, including study duration, number of measurements, number of patients, number of therapists, the level of randomization and the intraclass correlation between the levels. Determining the sample size that is needed for sufficient power to detect a treatment effect is more complicated than in other approaches, because there are different sample sizes at the different levels: the number of measurements per patient (level 1), the number of patients per therapist (level 2) and the number of therapists (level 3). Moreover, it is necessary to use plausible values of the variance components in the model to get a proper estimation of power, and these are usually unknown.

Although some literature (e.g. Raudenbusch, 1997; Snijders & Bosker, 1993) and software solutions (Bosker, Snijders, & Guldemond, 1996) are available to estimate the power for two-level models, there is little information about power analysis for three-level longitudinal models. The Optimal Design software developed by Raudenbush and colleagues (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2008) is the only program we are aware of with an option to compute power for three-level models,

and this functionality exists only for limited designs. In this paper we will demonstrate a method to compute power for three-level longitudinal models. With limited financial resources it is very valuable to know whether it is more efficient to collect more measurements per patient, a larger number of patients with fewer measurements or a larger number of therapists, with a smaller number of patients per therapist. First, we will discuss factors that influence power in more detail. Then, the influence of these factors on power will be illustrated using naturalistic data from an outpatient setting.

Factors influencing power

Intraclass-correlation

Ignoring a significant intraclass correlation (ICC) by using traditional regression models can result in substantial estimation bias for parameters and standard errors (Goldstein, 2003). The ICC is defined as the degree of resemblance between micro-units belonging to the same macro-level unit and can also be interpreted as the fraction of total variability that is due to nesting (Snijders & Bosker, 1999). In multilevel modeling, although nesting is taken into account by the model, the size of the ICC strongly influences the power to detect treatment and covariate effects. For three-level models intraclass correlations are calculated for levels two and three. There are multiple ways to compute the level-two intraclass correlations (see Hox, 2002; Snijders & Bosker, 1999).

The effect of clustering can be clearly demonstrated by the design effect. The design effect is the ratio of patients required for modeling nested data, relative to non-nested data and is defined as $1+(k-1)ICC$, with k being the number of patients per therapist, and the ICC defined as the proportion of the total variance accounted for by the therapist level (Donner & Klar, 2000). For example, with an ICC of 0.1 and four patients per therapist the design effect would equal 1.3, meaning that approximately 30% more patients are needed for sufficient power than in non-nested data. The higher the ICC, the larger the design effect and the lower is the power to detect a treatment effect. In this example a higher ICC means that there is a larger difference between therapists, and that the therapist variance explains part of the variance of the treatment effect, thus reducing the power to detect that effect.

One of the major challenges in a priori power analysis for multilevel models is the need for plausible values of the variance components, and these are usually unknown.. Using ICC values from the literature will provide an idea of the extent to which the design-effect will decrease power. As the ICC depends on the outcome measure, it is necessary to base the ICC estimates on studies using the same outcome measures. In educational and medical research, there is an increasing amount of literature

reporting intraclass correlations for different outcome measures and covariates (e.g. Adams et al., 2004; Campbell, Mollison, & Grimshaw, 2001; Hedges & Hedberg, 2007). In psychotherapy research, such papers are desperately needed in order to perform sensible a priori power analyses.

Level of randomization

Three-level longitudinal models offer a choice of which level should be used to randomize to treatment conditions. In terms of statistical precision and power it is usually best to randomize at the lowest level possible (Moerbeek, 2005). In the case of three-level longitudinal models in psychotherapy research this is the patient level (level two), since the first level consists of the repeated measurements. A common design in randomized controlled trials in psychotherapy research is that patients are randomly assigned to treatment conditions, and therapists are recruited to either provide the experimental or control treatment. Although patients are randomly assigned to treatment conditions, statistically the randomization is at the therapist level, since within a therapist all patients receive the same treatment. The reason that this design is so common, is that it prevents contamination of the treatment effect. Contamination occurs when members of the experimental treatment condition influence members of the control condition. If therapists provide both the control (for instance treatment as usual) and the experimental treatment, they might unconsciously use techniques from the experimental treatment for their patients in the control group, thus contaminating the treatment effect. In case of contamination, randomization at the therapist level, actually might lead to higher power to detect an effect (Moerbeek, 2005). Alternative designs like pseudo-cluster randomization (Borm, Melis, Teerenstra, & Peer, 2005) and the split-plot design (Reise & Duan, 2003) have been developed to help to handle contamination, but have not been widely used so far.

Sample size

In three-level longitudinal models there are different sample sizes at different levels: the number of measurements per patient (*m*), the number of patients per therapist (*k*) and the number of therapists (*J*). Several combinations of *m, k* and *J* can result in an identical power to detect an effect, but a minimum sample size at each level is necessary for accurate estimation of the estimates and standard errors. The most significant limitation on accurate estimation is usually the sample size at the highest level (Maas & Hox, 2005). For instance, when a relatively large portion of the variance is situated at the therapist level, and randomization takes place at this level, power does not increase much when more patients are included per therapist. Similarly, without

prolonging the duration of the study, the effect of additional measurements on power is limited (Moerbeek, 2008; Raudenbusch, 1988). However, having a larger number of measurements does provide more information about the progress of individual patients and makes it possible to fit models with a more complicated random part at the patient level (Snijders & Bosker, 1999).

Covariates

The effect of covariates on power is not always straightforward in multilevel models. The optimal sample size for each level can change when covariates are used, particularly when covariates are used at multiple levels. Determining the precise impact of a covariate on power a priori is complicated because it depends on a number of factors, including how much within and between variance a covariate accounts for (Reise & Duan, 2003). For example, should a covariate have a low correlation with the dependent variable, but a moderate to high correlation with other variables in the model (such as time or treatment condition) power is usually decreased by including the covariate in the model. However, if the covariate explains part of the variance that is unexplained by time or treatment condition, power is increased by its inclusion (Moerbeek, 2006).

Missing data and drop-out

One of the challenges of longitudinal research is the, almost inevitable, loss of data due to missingness or drop-out. Multilevel analysis is capable of handling most cases of missing data very well[1]. Missing data will, nonetheless, affect study power simply by the fact that fewer data points are available as a result. Moerbeek (2008) presented power plots for several drop-out patterns and showed that in studies with dropout, power especially decreases when the drop-outs are concentrated in the beginning of the study. Increasing the study duration can also have a negative effect on power, especially if the dropout is concentrated at the end of the study.

Data used as example

For a new study on the effect of providing feedback to therapists about their patients' progress, we sought to estimate how many therapists and patients would need to be included in order to obtain sufficient power to detect an effect. In the planned study, patients will be randomly assigned to either a feedback or control condition. In the feedback condition, therapists will receive charts by e-mail that indicate the patients progress on the Outcome Questionnaire (OQ-45; Lambert et al., 2003). Because the

variance components in the planned study are unknown, data from a routine outcome monitoring study were used to estimate the variance in the control group. The routine outcome monitoring (ROM) data consists of 1966 measurements of patient functioning (average of 3.22 measurements per patient), within 610 patients (average of 5.60 patients per therapist), who were treated by 109 therapists. Collection of the ROM data is an ongoing process and a portion of the patients in this dataset are still in treatment. Patients completed the OQ-45 at several moments during treatment. We used the log of the number of sessions as our time variable, a common choice in psychotherapy research that has been shown to linearize treatment progress. The intraclass correlation at the patient level[2] and therapist level were 0.75 and 0.18 respectively in the (empty) unconditional model. Several three-level multilevel models were fitted to the data and compared on the deviance values and Wald test for fixed effects. For explanatory purposes we used relatively simple models, even though there are numerous modeling options that might be relevant for this kind of data (e.g. anchoring, centering, transformations, additional covariates). Table 1 shows the results for a model with a random effect of (log) time at the patient level (Model A) and the effect of adding a covariate (gender) to the model (Model B). Adding a random effect of (log)time at level-two and three simultaneously demonstrated no significant improvement compared to Model A. Adding a random slope at the therapist level,

Table 1 Parameter estimates and standard errors for two multilevel models fitted to the routine outome monitoring data

	Parameter		*Model A*		*Model B*	
			Estimate	*SE*	*Estimate*	*SE*
Fixed effects						
Initial status, β_{0jk}	Intercept	δ_{000}	78.37	1.26	78.38	1.27
Rate of change, β_{1jk}	Slope	δ_{100}	-13.30	0.99	-10.56	1.64
	Gender	δ_{110}			-4.16	1.99
Variance components						
Level 1	Within-person		71.77	3.27	71.67	3.26
Level 2	Initial status		454.09	23.97	448.93	30.69
	Rate of change		187.80	29.57	183.72	29.25
Level 3	Initial status		48.61	19.40	49.53	19.61
	Rate of change					
Goodness of fit	Deviance			16123.5		16119.24
	Pseudo R^2					

SE = Standard error
Model A: random slope at the patient level
Model B: random slope at the patient level, gender as covariate

but not at the patient level, produced a significant improvement compared to a fixed effect of time, however, the model with a random slope at the patient level had a better goodness-of-fit.

Once we derived our best-fit model, we used the values from Model A as estimates for the variance and covariance components in the control condition of the planned study. The equations used to compute the power curves are described in Appendix A. A medium effect of 0.5 was expected for the interaction between time and treatment condition. For the power plots, the study duration was set at 21 weeks[3], with patients receiving one therapy session per week. The maximum number of therapists was set at 100. To demonstrate the effect of adding a covariate, we used the values from Model B. To show the effect of drop-out on power, we performed a simulation study using Splus. For each combination of number of therapists and drop-out pattern, 5000 data sets were simulated in which model parameters and standard errors were estimated. The observed power is reflected by the percentage of data sets in which the null hypothesis was not rejected; in other words, in percentage of data sets in which the effect of the feedback condition on the slope was not significant.

A priori power analysis using the example data

Figure 1 shows the estimated power curve for five measurements per patient, and two, four and eight patients per therapist. Since the slope variance at the therapist level did not significantly deviate from 0 in Model A, the therapist variance has no influence on the power to detect an effect (also see Appendix A), and consequently the figure applies to both randomization on the patient and the therapist level.

General guidelines state that a power level of 0.80 is considered adequate to detect treatment effects (Cohen, 2002). According to Figure 1, to obtain a power of 0.80 one should have either 22 therapists with eight patients each (176 patients in total), 43 therapists with four patients each (172 patients in total) or 85 therapists with two patients each (170 patients in total). In order to get unbiased estimations of the parameters and standard errors, a sufficient number of therapists is needed. In the simulation study by Maas & Hox (2005) for two levels, it was shown that with 30 groups the standard errors for the highest level were estimated about 15% too small; only with 50 groups were the chances of reliable estimation acceptable.

Optimal level of randomization

In the ROM data, the slope at the therapist level was relatively small and not significant. However, in some cases there might be a significant difference between therapists in how their patients progress over time. In that case, the level of randomization could

Figure 1 Estimated power for the planned study, no significant slope variance at the therapist level

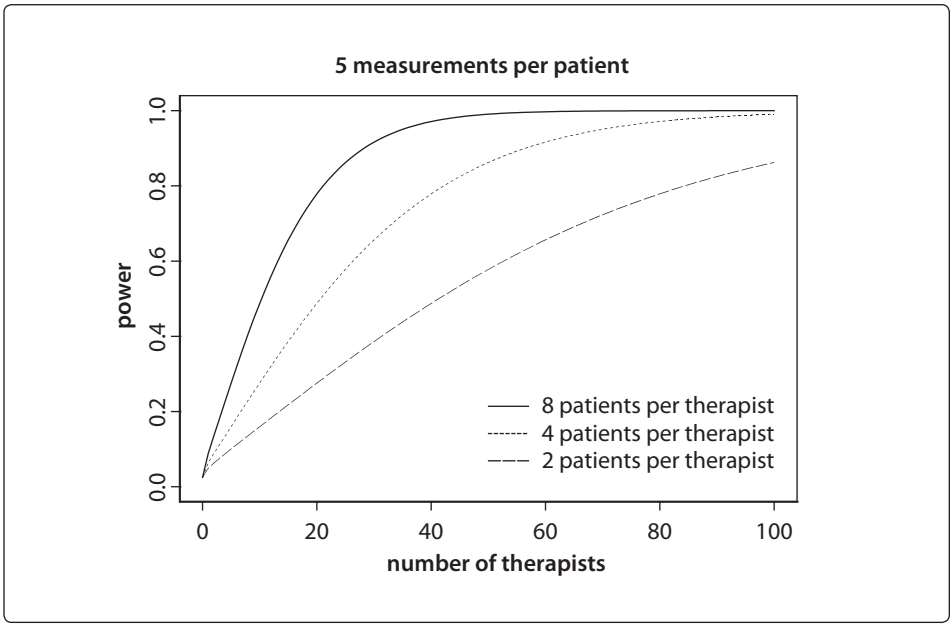
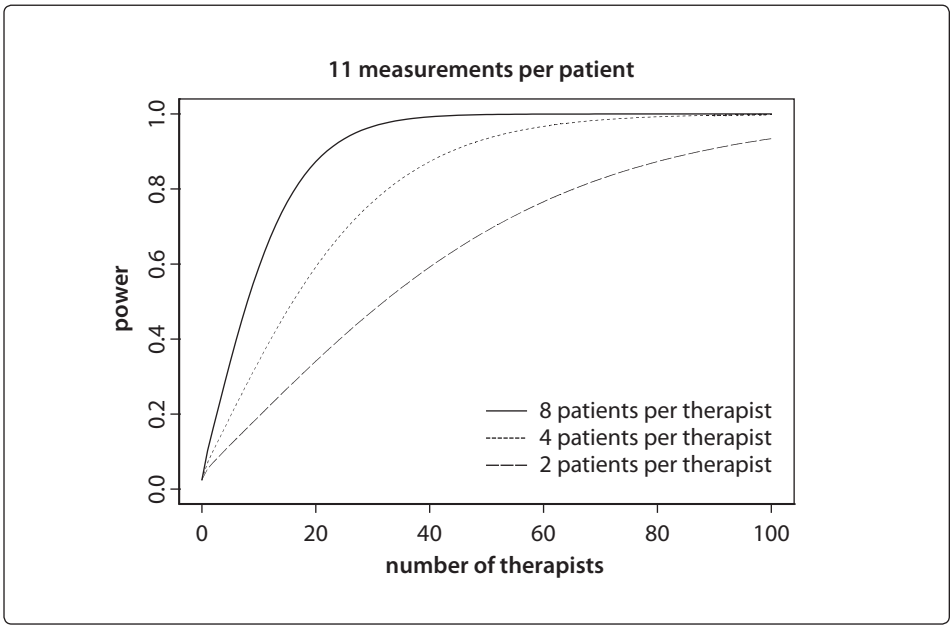


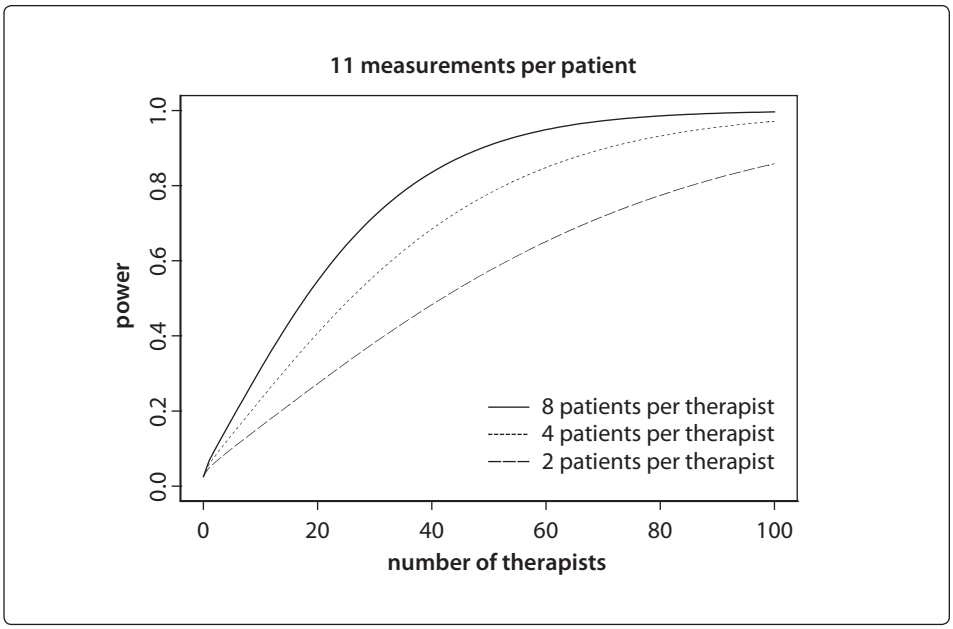
Figure 2a Estimated power, significant therapist slope variance, randomization at the patient level (level 2)



have a substantial impact on power. For this reason, we performed an additional power analysis with a significant slope at the therapist level. Based on the results from Lutz, Leon, Martinovich, Lyons & Stiles (2007), we assumed that beside the slope variance of 187.80 at the patient level that was found in Model A an additional variance of 35.77 (16% of the total slope variance) was situated at the therapist level. Figure 2a shows the power curve for this simulation, with randomization at the patient level, and Figure 2b shows the power curve for randomization at the therapist level, both for eight, four and two patients per therapist and 11 measurements per patient. When randomization takes place at the patient level (patients are randomly assigned to conditions, therapists are in both conditions) the samples required are 17 therapists with eight patients each, 33 therapists with four patients each, or 66 therapists with two patients each to reach a power level of 0.80. When randomization takes place at the therapist level (therapists are randomly assigned to conditions, all patients within a therapist are in the same condition) sample requirements are 37 therapists with eight patients each, 53 therapists with four patients each, or 86 therapists with two patients each.

Comparing Figures 2a and 2b also shows that adding patients per therapist has less effect on power when randomization takes place at the therapist level than at the patient level. Suppose we have 15 therapists with four patients each. When randomization occurs at the patient level, the power is 0.47. When the number of

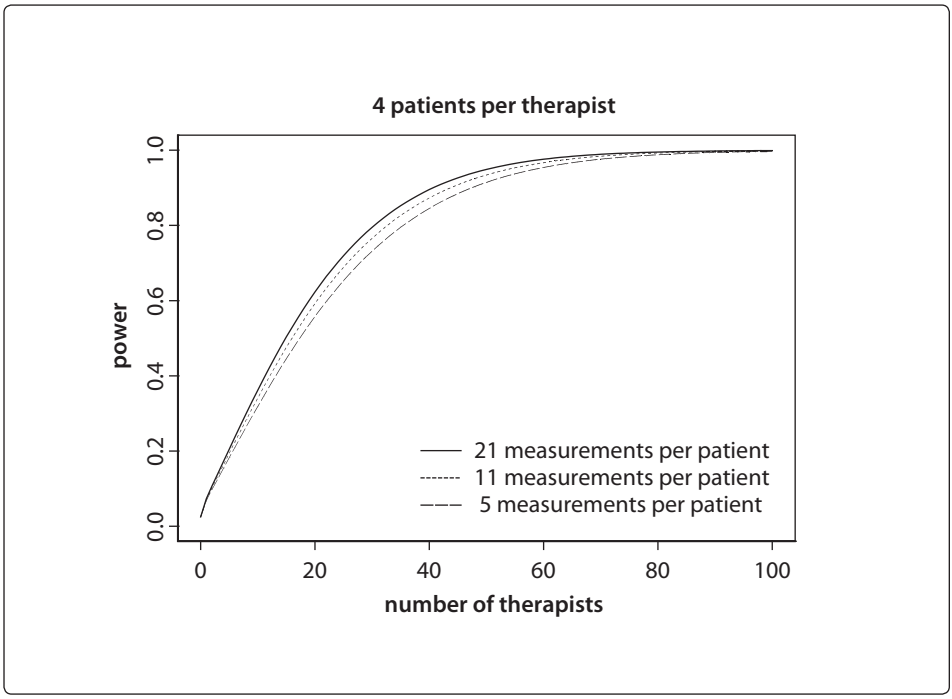
Figure 2b Estimated power, significant therapist slope variance, randomization at the therapist level (level 3)



patients per therapist is doubled, the resulting power is 0.77. When randomization takes place at the therapist level doubling the amount of patients per therapist increases the power from 0.32 to 0.44. However, doubling the amount of therapists in the study, rather than the amount of patients, results in a power of 0.56 and is, thus, far more effective.

Figures 2a and 2b demonstrate how randomization at the therapist level is less effective in terms of power than randomization at the patient level. Randomizing at the therapist level is only advised in case of contamination. In the planned study, contamination is not likely, as the studies performed by Lambert and colleagues (Harmon, Hawkins, Lambert, Slade, & Whipple, 2005; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert et al., 2002; Lambert et al., 2001; Whipple et al., 2003) show, therapists do not perform better with patients in a control condition over time. If contamination was an issue, effect sizes would have shrunk in each consecutive study as a result of the learning effect. In other words, if the feedback had taught therapists to treat all their patients more effectively, the effect of the feedback would have decreased with each consecutive study. Randomization at the patient level seems the best option for the study design in this case and, therefore, further power plots will only be presented for this situation.

Figure 3 Estimated power for a varying number of measurements per patient, randomization at the patient level



Number of measurement occasions

Earlier it was stated that the number of measurement occasions within the same time-period has little influence on power. To demonstrate this, power plots were produced for 5, 11 and 21 measurements. As can be seen in Figure 3 increasing the number of measurements per patient hardly has an effect on study power; respectively 36, 33 or 31 therapists are needed to obtain a power level of 0.80.

Adding a covariate

To demonstrate the effect of a covariate on power, gender was added as a slope predictor (see Figure 4). Although gender is a significant slope predictor, adding it to the model has little effect on power because it decreases the slope variance by only 2%. Stronger covariates might have a larger effect on power, provided that they don't increase variance on other levels. As is shown in Table 1, Model B, adding a covariate can increase the variance at other levels, in this case the initial status variance at level 3.

Figure 4 Estimated power with and without gender as a covariate, randomization at the patient level

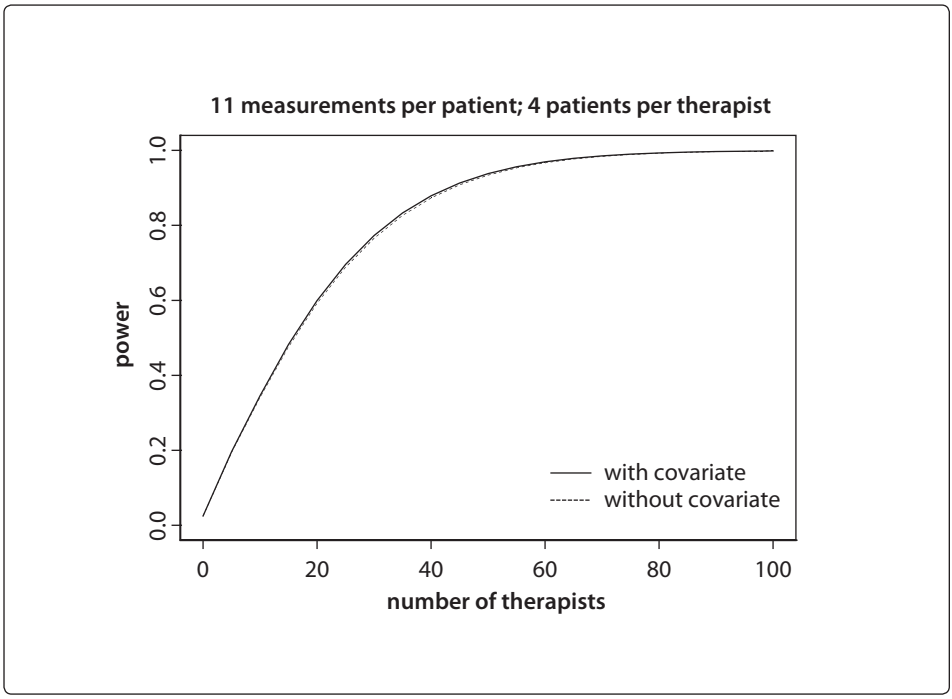


Figure 5a Simulated patterns of dropout (25% drop-out)

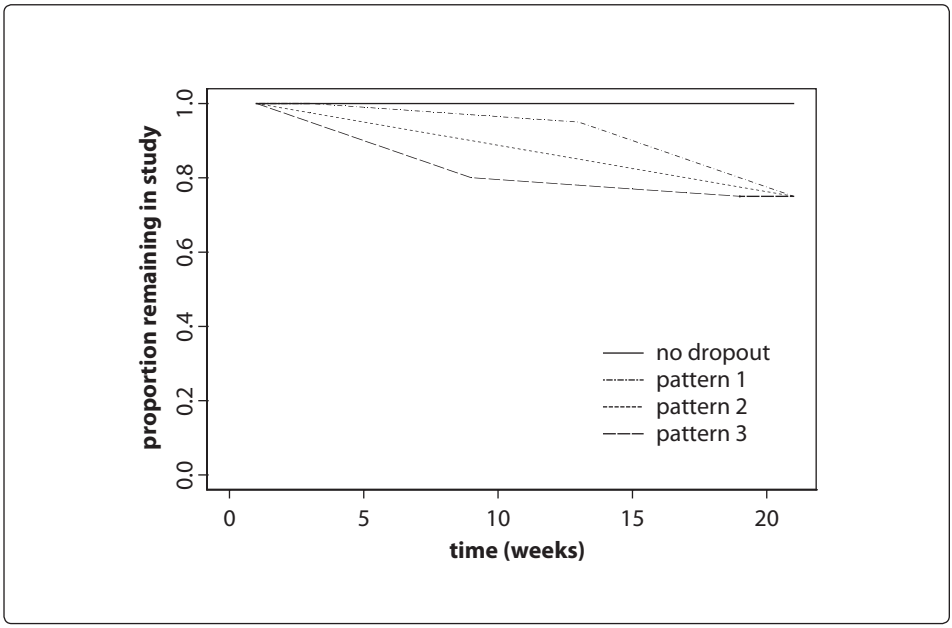
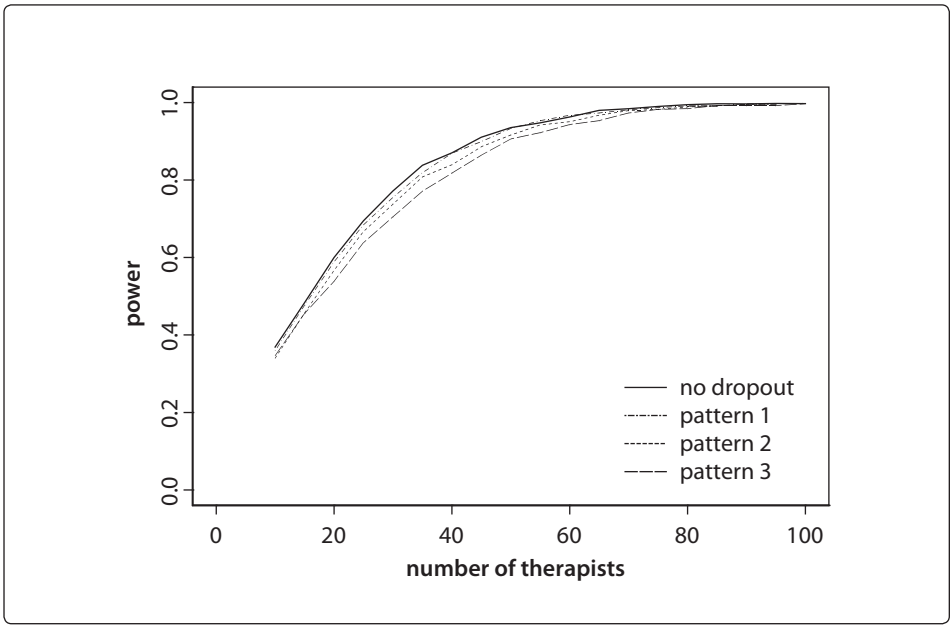


Figure 5b Estimated power for the simulated drop-out patterns, randomization at the patient level



The effect of study drop-out on power

Figure 5a shows the simulated patterns of drop-out for therapists with eleven measurements per patient and four patients per therapist. All patterns have a total drop-out of 25%. Pattern 1 has most of the drop-out at the end of the study; Pattern 2 has an equal drop-out throughout the study and Pattern 3 has most of the drop-out at the beginning of the study. Figure 5b shows the observed power for the simulated drop-out patterns. The effect of drop-out on the number of therapists needed for drop-out is limited. Having most of the drop-out at the beginning of the study seems to have the most substantial impact on power. Drop-out at the end of the study has the least impact on power. In a sample with no dropout, 31 therapists are needed. Drop-out at the end of the study results in a need for two additional therapists (33), equal drop-out throughout the study requires three additional therapists (34) and drop-out at the beginning in 6 additional therapists (37).

Discussion

The aim of this article was to perform an a priori power analysis for three-level longitudinal multilevel models and to demonstrate the effect of the level of randomization, samples size, covariates and drop-out on the power to detect a treatment effect. Results demonstrated that randomization at the patient level was more effective, in terms of power, than randomization at the therapist level. Increasing the number of patients was shown to be the best way to improve power when randomization takes place at the patient level. In the case of randomization at the therapist level, including more therapists in the study was more effective. The sample size at level one, the repeated measurements, did not have a strong effect on power. Furthermore, in our example adding gender as a covariate did not improve the power much. However, our covariate did not have a strong effect and other, more significant, covariates might have different effects on power. Drop-out also did not affect power substantially, although it did reduce power to some extent, especially when drop-out was concentrated at the beginning of the study. Besides power, it is necessary to have appropriate sample sizes at each level to ensure accurate estimation of parameters and standard errors. In some cases this may result in larger sample sizes than are necessary for sufficient power. In addition, in order to effectively distinguish between the slope variances at the patient and therapist level, there needs to be a sufficient level of patients per therapist.

Results indicate that, in three-level models, larger sample sizes are required than are common in general linear model approaches. It should be noted, however, that the intraclass correlation at level three in our example data was rather high.

This is consistent with naturalistic data, and one is less likely to find such values in a randomized controlled trial. Moreover, it is likely that the feedback condition in our planned study will reduce therapist variance and, by consequence, the level three intraclass correlation. In other naturalistic data we have found lower ICC values and, as a result, sufficient power with lower numbers. In such a case, power is no longer the main concern, and is secondary to estimation bias. For example, in our simulation, we found sufficient power using 17 therapists with eight patients each. However, by choosing that size sample, the model parameters and standard errors could be seriously biased, thus inflating the Type I error. This is specifically the case for the estimation of random effects; in fixed effects the standard errors are more robust. Maas and Hox (2005) found evidence for this in a simulated two-level model and although there are no simulations available for longitudinal three-level models, the problem of estimation bias likely applies to the highest level of the model and, thus, similar results can be expected for three-level models. Estimation bias in multilevel modeling is partly resolved by using non-parametric tests such as the likelihood ratio test instead of parametric tests, but this doesn't solve the problem completely.

The analyses that were performed in this article have some limitations. First, the modeling on which the analyses were based was kept deliberately simple. A random intercept, random slope model was selected, but fixed intercept or slope models could also have been used. For the time variable, a log-linear model was chosen, as this is the most frequently reported time variable in psychotherapy research, but other time variables are plausible as well (e.g. a combined linear and quadratic time variable). Another issue with the time variable is that calculating the dosage in sessions attended, rather than length of treatment, ignores frequency effects that would be found in, for example, twice weekly versus once-weekly psychotherapy. In addition, psychotherapy studies usually have several relevant predictor variables, rather than just one, and there would be predictors for both the slopes and the intercepts at the patient and therapist level. Since intercept predictors do not influence the power to detect a treatment effect, they were not included in the model. An additional limitation of this article is that it only describes power for situations with two conditions. Many clinical trials have three conditions, comparing two experimental conditions and one control condition. In such a case, the model becomes more complex quickly, because an additional (dummy) variable has to be added. One has to know in advance how the three conditions will affect slope and intercept variances. Lastly, although this article has provided examples how several relevant factors such as covariates and drop-out influence power, it has only addressed these phenomenon one at a time, whereas in practice multiple such factors frequently apply at the same time. However, these results offer a good indication of how each factor influences power in these models and could help in making decisions about study designs in the future.

Although a priori power analysis for multilevel models is a complex undertaking, these results indicate that such modeling is not only possible, but practical. Traditional power analyses for linear models does not distinguish between sample sizes at different levels and can lead to an underestimation of the number of cases that are needed. Therefore, although a growing literature on multi-level power analysis exists, further exploration is warranted. In particular, in order to be able to perform a priori power analyses for clinical trials, there is a strong need for more articles on variance components and intraclass correlations in different types of patients, treatments and outcome measures.

Since the special section in this journal on therapist effects, the discussion about whether therapist effects exist has become mixed up with the discussion on what models should be used to investigate them. Wampold & Bolt (2006) have stated that longitudinal models may increase patient variability and thereby reduce therapist effects. While this is true, patient variance in treatment progress is an integral part to the kind of data we collect in psychotherapy studies, as patients differ in their treatment course. By ignoring that variance one might be able to better detect therapist effects, but does that mean it is better? We do not claim to have the answer to that question, and would like to state that, in our opinion, it is more a matter of what the focus of the study is, than one method being better than the other. Longitudinal models do have some disadvantages, for instance, in the case of differential treatment length, longer treatments will have more impact on the model as they have more measurements than shorter treatments. Another issue is that longitudinal models do not assess whether treatment is successful or not, but neither do end-of-functioning models.

Irrespective of the method or model used, the evidence for therapist effects is limited so far (Crits-Christoph & Gallop, 2006). The lack of evidence for therapist effects in clinical trials seems to have two main causes: the number of trials that take the therapist level into account is small and the trials that do include therapist effects often have small sample sizes that are too small for unbiased estimation or sufficient power. In order to get more information on the existence of therapist effects in clinical trials, the therapist level should, at least, be registered and reported, regardless of significance. In addition, predictor variables at the third level should be included in more studies, to explain what factors may contribute to therapist effects.

Acknowledgements

Mirjam Moerbeek's research was funded by the Netherlands Organisation for Scientific Research (NWO), grant number 451-02-118. We would like to thank Sam Nordberg for the corrections he made to improve the English.

Notes

I Multilevel analysis can handle data that is Missing Completely At Random (MCAR) or Missing At Random (MAR), but not data that is Missing Not At Random (MNAR).

II The patient level variance has been calculated according to Davis & Scott, 1995 cited in Hox, 2002 (p.32)

III A study duration of 21 weeks results in equally spaced, round session numbers for the measurement occasions, as well as pre and post test. For 5 measurements the occasions are set at session 1, 6, 11, 16 and 21; for 11 measurements at sessions 1, 3, 5,…21; and for 21 measurements at each session.

Chapter 4

Risk models for negative treatment outcomes in psychiatric outpatients: predicting end state functioning and rate of change using classification and regression trees (CART) and multilevel modeling

De Jong, K., Nugter, M.A., Ninaber, C., Lutz, W.,
Van Ginkel, J.R., Heiser, W.J. & Spinhoven, P.

Manuscript submitted for publication.

Abstract

Objective. Risk models that aimed to identify consistent predictors for negative outcomes have encountered several challenges. This study uses state of the art statistical techniques to handle these problems, by using multilevel analysis combined with multiple imputation to predict the rate of change and classification and regression tree (CART) analysis to predict end state functioning.

Method. A naturalistic sample of 1540 outpatients (63% female; age range 17-67 years, *M* = 37.5, *SD* = 11.7) was collected in three mental health care organizations in the Netherlands. Patients completed the Outcome Questionnaire (OQ-45; Lambert et al., 2004) regularly during treatment. In addition, several potential predictor variables were collected.

Results. Initial severity, educational level, expectancies, Global Assessment of Functioning (GAF) and the working alliance were significant predictors for end state functioning. In predicting rate of change, the same predictors were found, except for educational level and expectancies. In addition, previous treatment, comorbidity and having a personality disorder as main diagnosis were significant predictors for rate of change as well.

Conclusions. Although there was overlap in predictors of negative outcomes with regard to end state functioning and rate of change as outcome variables, both analyses provide different information. By combining the prediction models, patients that may need to be monitored more closely during treatment can be identified so that negative outcomes may be prevented. By using CART and multilevel analysis combined with multiple imputation substantially more data could be used in analysis than would otherwise have been the case, thus reducing the selection bias and improving generalization.

Introduction

Risk models that aim to predict the future course and outcome of disease processes are common in medical and health research. Good risk models are valuable for a wide variety of purposes, including policy making, adjusting for differences in patient case-mix between institutions, and assisting patients and clinicians to make informed decisions about treatment (Ambler, Omar & Royston, 2007). In medical situations, knowing the risk factors for negative treatment outcomes might mean the difference between life and death. Although psychotherapy is usually not involved in life and death situations, negative treatment outcomes can have a large impact on patients' quality of life and potentially constitute an increased risk of long-term psychiatric complaints and higher costs for mental health care. Hansen, Lambert and Forman (2002) have shown that in naturalistic settings a lack of treatment success is common: 3-14% of patients deteriorate and 45-60% show no clinically significant change during treatment. By comparison, in clinical trials results are much better: 67% of patients improve significantly. Similar results were found by Barkham et al. (2008), who showed that approximately 18% more patients were clinically significantly improved and effect sizes were more than twice as large in randomized trials than in practice-based studies. This suggests that there is much room for improvement in clinical practice. Although not all patients are likely to achieve treatment success and some might be on a progressive decline that cannot be stopped, evidence suggests that at least some patients might worsen as a result of therapy (Lambert & Ogles, 2004). Therefore, knowing risk factors for negative treatment outcomes in psychotherapy can be very valuable for treatment selection.

In psychotherapy research, risk models are best known from the line of research that is referred to as *patient-focused research* and health services research such as quality assurance programs. Such research aims to prevent negative treatment outcomes by making predictions about the patient's progress using variables that have previously been associated with positive or negative treatment outcomes (e.g. Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lueger et al., 2001; Lutz, 2003). These models usually attempt to forecast the course of treatment or rate of change, rather than end state functioning, and are often referred to as expected treatment recovery (ETR) curves. Most models use multilevel modeling or similar techniques to make these predictions. Although these models are very valuable, it has been a challenge to identify reliable predictors of change beyond initial severity. Some authors have found additional predictors, including patients' expectancies of outcome and the Global Assessment of Functioning score, but in these studies initial severity still explained the highest proportion of outcome variance, compared with all other predictors (Lutz et al., 2005; Lutz, Lowry, Kopta, Einstein, & Howard, 2001; Lutz, Martinovich, & Howard, 1999).

Moreover, most factors have not been consistently replicated by others. For outcome in terms of functioning at the end of treatment, initial severity is an important predictor too. In the review by Clarkin & Levy (2004), a high initial severity of symptoms was related with poor treatment outcomes, especially in depression and addiction populations. The results on other client variables were less consistent: comorbid personality disorders for instance could have a positive or negative effect on outcome and client demographics were usually not consistently found as relevant predictors (Clarkin & Levy, 2004).

One reason for not finding many predictors for end state functioning and rate of change might be that we oversimplify the relationship between the predictor variable and outcome. Most prediction models suppose a linear relationship between the predictor and outcome, whereas the true relation might be much more complex. Kendler (2008) states that we need to start looking at more complex interactions in explanatory models for psychiatric illness – models that consider predictors at different levels: micro and macro, within and outside the individual – to better understand what risk factors are relevant in psychiatry. Assuming simple linear relationships between a predictor and outcome may be misleading. Take for example a predictor like the working alliance. The working alliance is one of the more robust predictors of outcome that has been identified (e.g. Horvath & Symonds, 1991; Klein et al., 2003; Martin, Garske, & Davis, 2000). It is often assumed that a strong alliance leads to better outcomes, yet not all patients with a high working alliance have good outcomes. Suppose that the true relationship would only hold for those patients with average to low severity of symptoms and not for patients with a high level of symptom distress. Standard regression models, would only find working alliance to be a predictor and might miss the interaction.

Another challenge for developing accurate risk models is ensuring sufficiently complete data. Most risk models for health outcomes are estimated using data routinely collected in clinical practice. The advantage of that approach is that there is usually a large dataset available, with high external validity. The disadvantage is that those datasets typically have many missing values. A review of more than 800 articles published in three leading personality journals showed that almost half of the articles reported missing data problems (Van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010). It is not unusual for some important predictors to be missing for over 50% of patients (e.g. Ambler et al., 2005). Moreover, in many studies patients have missing data on several variables simultaneously. Combinations of missing data on different predictor variables may result in percentages of complete data as low as 32% (Ambler, Omar, & Royston, 2007). Although missing values are common, information on the percentage of missing data and how missing data are handled statistically is often not reported. A survey of 46 papers in a counseling psychology journal showed that only a little

over one third of the papers reported missing data percentages (Schlomer, Bauman, & Card, 2010). A review of 100 articles using longitudinal data in three leading journals in developmental psychology showed that 82% of the studies that did report having missing data, used missing data techniques that are problematic (Jelicic, Phelps, & Lerner, 2009). In addition, anecdotes from other psychotherapy researchers tell us that it is not uncommon for more than half of the data to be "cleaned up" prior to final analysis (see Hatfield, McCullough, Frantz, & Krieger, 2010 for an example) although this is seldom reported in the published articles. Complete case analysis can result in substantially smaller sample sizes, biased regression coefficients and reduced reliability for predicting future observations as a result (Ambler, Omar & Royston, 2007). Moreover, researchers have demonstrated that serious violations of statistical assumptions occur when missing data are ignored (e.g. Allison, 2003; Graham & Hofer, 2000; Wothke, 2000).

Several statistical techniques have been developed that handle missing data problems well, including repeated measures multilevel analysis (Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002; Singer & Willet, 2003) and classification and regression trees (CART; Breiman, 1984). Multilevel analysis handles missing data on the dependent variable by estimating an individual change trajectory that depends on the observed variables for each person and as a result can handle missing data on the dependent variable very well, but it does need complete values on predictor variables. CART is a regression based datadriven technique that automatically searches for moderators in the data and calculates optimal splits for predictor variables. The missing data are handled by the use of surrogate splits. These type of split points are created as backup for the main split, meaning that if the split variable being evaluated is missing, a surrogate predictor is used (Breiman, 1984). Contrary to most techniques it can handle missings on the predictor variables very well, but needs complete cases on the dependent variable. CART has the additional advantage of being able to select important predictors automatically, especially when there are many variables (categorical and/or interval), and has the ability to uncover complex interaction effects between them. Furthermore, the method is robust for non-normally distributed data (Briand, Ducharme, Parache, & Mercat-Rommens, 2009). The resulting trees are easy interpretable by lay persons and can be used to create decision rules, which can be generalized to a non-research setting (Lewis, 2000). CART and multilevel analysis address different aspects of outcome: multilevel analysis assesses factors that are related to the rate of change, whereas CART assesses factors related to end of treatment functioning.

Another approach to handling missing data is multiple imputation. Multiple imputation can be used to address missing data in both independent and dependent variables and restores the dataset to the full size. Many authors have illustrated that

multiple imputation gives better results in statistical analyses than listwise deletion (e.g. Ambler, et al., 2007; Little & Rubin, 2002; Rubin, 1987; Schafer, 1997; Schlomer, et al., 2010). It decreases estimation bias in both the parameters and the standard errors and when the imputation model is correct and the data are missing at random, the precision of the parameter estimates will approach the precision that would have been achieved with complete data (Huang et al., 2009). However, clinical researchers seem to be reluctant to use it (Jelicic, et al., 2009). Although multiple imputation is a very useful method to deal with missing data and has been rapidly developed for many types of data, it still has trouble handling unbalanced designs like those that are common in naturalistic settings in clinical psychology. For example, patients' progress is often measured multiple times during the course of treatment, yet the frequency of sessions and assessments may differ considerately among patients - one patient might have 5 sessions of treatment with one measurement missing whereas another patient has 15 sessions with 3 measurements missing. Even with complete data the number of measurements would differ between patients. In the currently available software packages, the full data matrix is filled up, rather than just the true missing values. In this situation, multiple imputation can be used to impute the predictor variables, but not the outcome variable.

In summary, developing prediction model for negative treatment outcomes is important, since many patients in clinical practice do not experience sufficient change. Yet, handling the missing data that are typical in the naturalistic data used for these models poses a challenge. Applying techniques that are more flexible in handling missing data seem a good solution, but these methods have their limitations as well. In the current article, we will use a naturalistic dataset collected in three community outpatient facilities in the Netherlands to predict patients at risk for negative treatment outcomes. Data on the predictor variables will be multiply imputed and multilevel analysis will be used to predict factors related to the rate of change. Factors directly related to negative outcomes will be assessed using CART on the original data. The predictor variables in the multilevel analysis and CART analysis will be identical and include clinical variables (e.g. initial severity, diagnosis, duration of complaints), demographic variables (e.g. sex, age, educational level) and process variables (expectancies, working alliance). Differences in the results and the performance of both models will be compared.

Method

Participants

Data were collected between June 2006 and June 2009 in eight treatment departments of three mental health care organizations in the Netherlands. Subjects were outpatients who were seen for psychological or psychiatric treatment. Data were collected as part of routine care, but patients were offered the option to refuse participation. The research proposal was evaluated by the local ethical committees of the participating institutions. There were 4447 patients in the original sample, including 1689 males (38%) , 2752 females (62%) and 6 persons with unknown gender. The age of the patients ranged from 17 to 71, with a mean of 37.7 years (*SD* = 11.8). Patients with psychotic disorders, mental retardation or in a current crisis at the time of referral (*n* = 92), patients who received non-verbal treatments (*n* = 151), patients who did not have a sufficient level of understanding of Dutch (*n* = 121) and patients who did not receive more than one treatment session (*n* = 148) or were unable to participate in the study for other reasons (*n* = 91) were excluded from the study. A total of 503 patients actively refused to participate in the study, and an additional 1801 patients completed less than two questionnaires during treatment - 773 of which never completed a single questionnaire. The remaining sample of 1540 consisted of 561 men (36%) and 976 women (63%) and 3 people of unknown gender. The age of the subjects ranged from 17 to 67 with a mean age of 37.5 (*SD* = 11.7) (see Table 1).

A subset of 541 patients was used for the complete case analysis in the multilevel model, to demonstrate the effect of the multiple imputation on the analysis. Since only 19% of patients (*n* = 295) had complete values on all predictor variables, an optimal set of predictor variables was selected in such a way that the most relevant predictor variables could be included, without losing too many other cases (see Table 2).

Table 1 shows the baseline characteristics of the sample that met inclusion criteria, the final selection and the complete case sample. There were no significant differences found between the samples in sex, age, main diagnosis for treatment following the *Diagnostic and Statistical Manual of Mental Disorders IV* (DSM-IV) and Outcome Questionnaire – 45 (OQ-45; Lambert et al., 2004) scores at intake. The most common main diagnosis in our sample was mood disorder, followed by anxiety disorder and adjustment disorder.

Instruments

Outcome Questionnaire-45 item version (OQ-45)

The Dutch translation of the Outcome Questionnaire-45 item version (OQ-45; Lambert et al., 2004) was used to measure patient progress during treatment. The OQ-45 is

Table 1 Patient characteristics of the sample meeting inclusion criteria, the sample selected for main analyses and sample selected for the complete case analysis

	Sample meeting inclusion criteria (n = 3844)	*Selected for main analyses (n = 1540)*	*Complete case analysis (n = 541)*
Sex			
- Male	1419 (37%)	561 (36%)	197 (36%)
- Female	2419 (63%)	976 (63%)	344 (64%)
- Unknown	6 (0.2%)	3 (0.2%)	
Age	*M* = 37.6 *SD* = 11.8	*M* = 37.5 *SD* = 11.7	*M* = 38.5 *SD* = 11.6
Main DSM-IV diagnosis			
- Mood disorder	1149 (30%)	466 (30%)	170 (31%)
- Anxiety disorder	658 (17%)	302 (20%)	106 (20%)
- Adjustment disorder	580 (15%)	269 (18%)	93 (17%)
- Disorders usually first diagnosed in childhood	215 (6%)	78 (5%)	24 (4%)
- Personality disorder (Axis II)	236 (6%)	102 (6%)	33 (6%)
- Cognitive disorder	136 (4%)	71 (5%)	47 (9%)
- Substance related disorders	86 (2%)	19 (1%)	8 (2%)
- Other	486 (13%)	193 (13%)	59 (11%)
- No DSM-IV Axis I or II diagnosis	8 (0.2%)	6 (0.4%)	1 (0.2%)
- Unknown or missing	290 (8%)	34 (2%)	-
OQ-45 intake score	M = 77.5[a] SD = 23.9	M = 77.4[b] SD = 23.0	M = 76.6 SD = 23.0

Note. DSM-IV = Diagnostic and Statistical Manual of Mental Disorders, 4th Edition; OQ-45 = Outcome Questionnaire-45 item version
[a] n = 2452, [b] n = 1419

a self-report instrument and has 45 items, 9 of which are reversed, asking how the respondent has felt over the last week on a 5 point rating scale, ranging from 0 (never) to 4 (almost always). The OQ-45 consists of three subscales that assess different domains of client functioning: Symptom Distress, Interpersonal Relations and Social Role. The Symptom Distress domain consists of 25 items relating to psychological symptoms that are common in highly prevalent mental disorders. The Interpersonal Relations domain consists of 9 items that assess the functioning of the patient in interpersonal relationships, and the Social Role domain assesses the patients functioning in social roles, such as work and school. The Dutch OQ-45 has satisfactory psychometric properties. The internal consistency for the Total score ranges between 0.92 and 0.96 in university, community, patients and community and patients combined samples. For the subscales the international consistency is 0.90-0.95 for the Symptom Distress (SD) scale, 0.74-0.84 for the Interpersonal Relations (IR) subscale and 0.53-0.72 for the Social Role (SR) subscale (De Jong, Nugter, Lambert, & Burlingame, 2009).

Working Alliance Inventory (WAI)

The therapeutic relationship between therapist and patient was assessed using the Working Alliance Inventory Client's Form (WAI; Horvath & Greenberg, 1989). The Dutch version of the WAI is called the Werkalliantie Vragenlijst (WAV; Vervaeke & Vertommen, 1996) and consists of 36 items that are scored on a 5-point rating scale, ranging from 1 (never) to 5 (always). An example of an item from the WAI is 'I believe _____ (therapist's name) likes me'. The WAI has three subscales that consist of 12 items each: The Bond, Task and Goal subscales. The Bond subscale assesses the therapeutic bond between the patient and therapist, the Goal subscale measures the level of agreement in therapy goals between patient and therapist, and the Task subscale assesses the level of agreement between therapist and patient on who has to do what in the treatment. The Dutch version of the WAI had internal consistencies for the Bond, Task and Goal subscales of 0.85, 0.88 and 0.88 respectively (Vervaeke & Vertommen, 1993).

Treatment Credibility Questionnaire (TCQ)

The patient's expectations of the treatment results were measured with a questionnaire derived from the Treatment Credibility Questionnaire (TCQ; Borkovec & Nau, 1972). Our version was based on the adaptation by Addis et al. (2004), with one additional item ("How much improvement in your symptoms do you think will occur") from the new version of the TCQ, the Credibility Expectancy Questionnaire (Devilly & Borkovec, 2000). The TCQ version that was used in this study consisted of 7 items that are scored on a 7-point rating scale, ranging from 1 (not at all) to 7 (extremely). The TCQ consists of two factors, Expectancies and Credibility. The Expectancies subscale consists of 4 items and assesses patients' expectations of therapy outcome. The Credibility subscale consists of 3 items and measures the degree to which the patient thinks the therapy is credible. The version that was used in this study has an internal consistency of 0.89 for the Expectancies subscale and 0.84 for the Credibility subscale (Hüpscher, 2007).

Patient characteristics

A variety of patient characteristics were extracted from the electronic registration systems of the participating mental health care institutions. Patient characteristics included demographic variables such as age, gender, education, and clinical variables such as DSM-IV diagnosis and prior treatment (see Table 2 for the full list). The information that was retrievable differed between the participating institutions. In addition, information from intake forms and electronic patient files was used in order to complete missing data when possible.

Procedure

Patients were informed about the study through a letter, before the intake. Participation in the study was on a voluntary basis and patients could object to participate by filling out the enclosed rejection form, but were automatically included if they did not reject. Patients were asked to report to the reception desk fifteen minutes before their treatment session started. The receptionist provided them with an OQ-45. The therapist handed out the TCQ and WAI to the patient after session 2. The patient could hand in the completed forms the next session at the reception desk. Patients were asked to complete the OQ-45 at intake, the first five sessions of treatment and subsequently every fifth session of treatment (the tenth, fifteenth, etc.), for a maximum period of one year. If patients were still in treatment after one year, the final measure was administered at that time; this applied to 445 patients (n = 1494; 30%).

Analysis

Missing Data

In the final sample, only 295 (19%) of the 1540 patients had completely observed responses on all the relevant predictor variables. The amount of missing values ranged between 0 and 20, with a mean of 5.3 (SD = 3.8). Missing data on the predictor variables were handled using Multiple Imputation using Chained Equations (MICE; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). In general, multiple imputation works as follows: Missing data are estimated multiple times using one or more statistical models. The procedure then generates plausible random values for the missing data that resemble the observed data as much as possible. By estimating the missing data multiple times several complete versions of the incomplete dataset are created, which are analyzed by standard statistical procedures. The results of these analyses are then pooled into one final analysis. Multiple imputation was carried out using PASW 17.0 (SPSS, 2009).

The MICE procedure uses linear regression to estimate the missing values on continuous variables, using the other variables as predictors, and logistic regression to estimate the missing values on categorical variables. Van Buuren, Boshuizen & Knook (1999) recommend using a maximum of 25 variables. Thus, a selection of variables was used which were necessary for carrying out the statistical analyses of interest. The variables that were used in the imputation procedure are given in Table 2. Based on recommendations by Graham (Graham, 2009; Graham, Olchowski, & Gilreath, 2007) data were imputed 20 times and the results of the 20 imputed databases were combined using Rubin's (1987) rules for multiple imputation. PASW 17.0 (SPSS, 2008) automatically combines the results of multiply imputed datasets into one pooled analysis.

Definition of negative outcome

The OQ-45 was used to define treatment outcome, based on the score at the end of treatment, or, if that score was missing, the last available score for that patient. Negative treatment outcome was defined as a patient scoring in the clinical range (55 or higher on the OQ-45) at the end of treatment and having experienced *no change* (less than 14 points decrease in score on the OQ-45) or *deterioration* (more than 14 points increase in score) according to the criteria for reliable change within the concept of clinical significance by Jacobson & Truax (1991). This definition is slightly different from that used by others. Usually negative outcomes are defined as deterioration or no change, regardless of the level of functioning at the end of treatment. We think that people who are functioning in the normal range at the end of treatment should not be considered as having experienced a negative outcome, so we excluded this group from the negative outcomes group. Based on the definition that we used, 727 (49%) patients in our sample had negative treatment outcomes, of which 130 patients deteriorated and 597 showed no (reliable) change.

Classification And Regression Trees (CART)

The Classification And Regression Trees (CART) method to acquire a valid model can be broken down in a couple of steps. The CART algorithm searches through the predictor values for a split point. This way the data are split up in subsets, also called nodes, and the homogeneity of the outcome variable in the new subgroups has increased. The best split is selected, based on the criterion that there is an optimal division of the outcome variable. This partitioning of data is then repeated for the new (sub)sets. In other words, we start with the full dataset, which is split up based on an optimal cut point value on a particular predictor and then the same procedure is applied to the new sets, also called child nodes. Repeating this process many times will lead to a situation where the number of subjects in a child node becomes one or they all have the same outcome. In case of the latter, the partitioning process can be stopped. It can be reasoned that the large model constructed may be poorly generalizable, in other words an overfitted model, because it is tailored for a particular dataset. To overcome this problem Breiman (1984) proposed a regularization method, which aims to find the optimal trade-off between the size of a tree and its predictive power. We start with a very large tree, which has overfitted the data. Each branch causes some amount of homogeneity gain in their end nodes. But this reduction needs to be viewed in relation to the size of the branch, for example a branch early in the tree probable leads to a larger reduction than one almost at the end. This ratio, often referred to as the cost-complexity ratio, is used to select the branch which benefits the model the least and pruned (removed) out of the tree. This pruning procedure is then sequentially repeated and the optimal size of the final tree is determined based on cross-validation.

Table 2 Predictor variables prior to multiple imputation (n = 1540) and correlations with change and the last available OQ-45 score (end).

	r (change)	*r* (end)	% missing	In complete case analysis	Values
Sex	-0.03	0.06	0%	Yes	See Table 1
Age	-0.04	0.04	0%	Yes	See Table 1
Main DSM-IV diagnosis category			2%	Yes	See Table 1
- Mood disorder	-0.12	0.25			
- Anxiety disorder	0.06	-0.11			
- Adjustment disorder	-0.05	-0.04			
- Disorders usually first diagnosed in childhood	0.04	-0.05			
- Personality disorder	0.06	0.03			
- Cognitive disorder	0.05	0.02			
- Substance related disorders	0.01	0.08			
- Other	0.02	-0.16			
Educational level	-0.06	-0.03	36%	No	31% Low, 32% middle, 33% high, 4% other
Having a paid job	-0.02	-0.15	32%	No	64% Yes
Previous treatment	0.04	0.12	31%	No	55% Yes
Using psychiatric medication at intake*	-0.03	0.23	34%	No	49% Yes
Duration of complaints	0.05	0.01	34%	No	52% Longer than one year
Diagnosis on Axis I	-0.05	0.07	2%	Yes	98% Yes
Comorbidity on Axis I	0.01	0.18	2%	Yes	34% Yes
Diagnosis on Axis II	0.03	0.10	2%	Yes	17% Yes
Comorbidity Axis I and II	0.02	0.12	2%	Yes	16% Yes
Problems with primary support group	-0.05	0.11	13%	No	46% Yes
Problems related to social environment	0.01	0.06	13%	No	18% Yes
Occupational problems	-0.05	0.10	13%	No	33% Yes
Legal problems / crime	0.00	-0.09	13%	No	1% Yes
GAF score (Axis V)	0.01	-0.22	6%	Yes	M=59.6, SD=8.6
OQ-45 SD subscale	-0.33	0.89	8%	Yes	M=48.3, SD=15.3
OQ-45 IR subscale	-0.22	0.71	8%	Yes	M=16.1, SD=6.5
OQ-45 SR subscale	-0.26	0.64	10%	Yes	M=13.0, SD=5.2
TCQ Expectancies subscale	-0.05	-0.13	47%	Yes	M=21.2, SD=4.2
TCQ Credibility subscale	-0.03	-0.11	47%	Yes	M=15.1, SD=3.2
WAI Bond subscale	-0.03	-0.15	56%	Yes	M=49.6, SD=6.4
WAI Task subscale	-0.02	-0.24	56%	Yes	M=48.7, SD=6.8
WAI Goal subscale	-0.02	-0.23	56%	Yes	M=48.1, SD=6.7

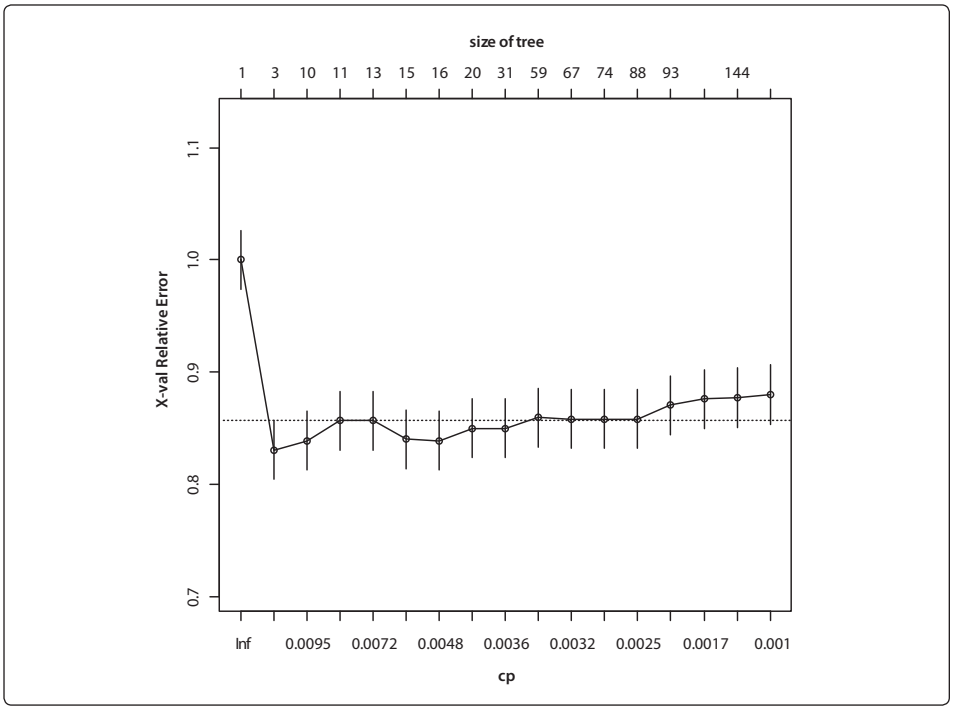
Note. Correlations between the predictor variables and change and the last available measurement of the OQ-45 (end) are pooled estimates, based on the imputed dataset. * Psychiatric medication had already been prescribed by the general practitioner for these patients

For the building of the initial models based on the CART method, the R package Rpart (Therneau & Atkinson, 1997) was chosen, because it closely follows the work and propositions of Breiman (1984). To acquire a large tree, which could later on be pruned back, the following model parameters were used. The maximum depth possible for a tree was 30 layers. The minimum number of cases in a terminal node was set to 10 cases. The complexity parameter was set to 0.001. The Gini splitting criterion was used because it in general is more favorable (Breiman, 1984). The optimal classification threshold was set at 0.50, based on ROC curve analyses for the two final models.

Multilevel Analysis

Two multilevel analyses were performed, on the complete case sample and on the multiply imputed datasets. Both models were two-level random intercept random slope multilevel models, using maximum likelihood estimation and with an unstructured covariance structure. The analyses were performed in PASW 17.0 (SPSS, 2009). Time was set as the logarithm of the session number. A backwards procedure was applied, starting with a full model and removing non-significant predictors (based on the Wald test for fixed effects) one by one until a parsimonious model was reached that was not significantly worse than the full model.

Figure 1 Cost-complexity versus relative error per tree size (number of nodes)



Results

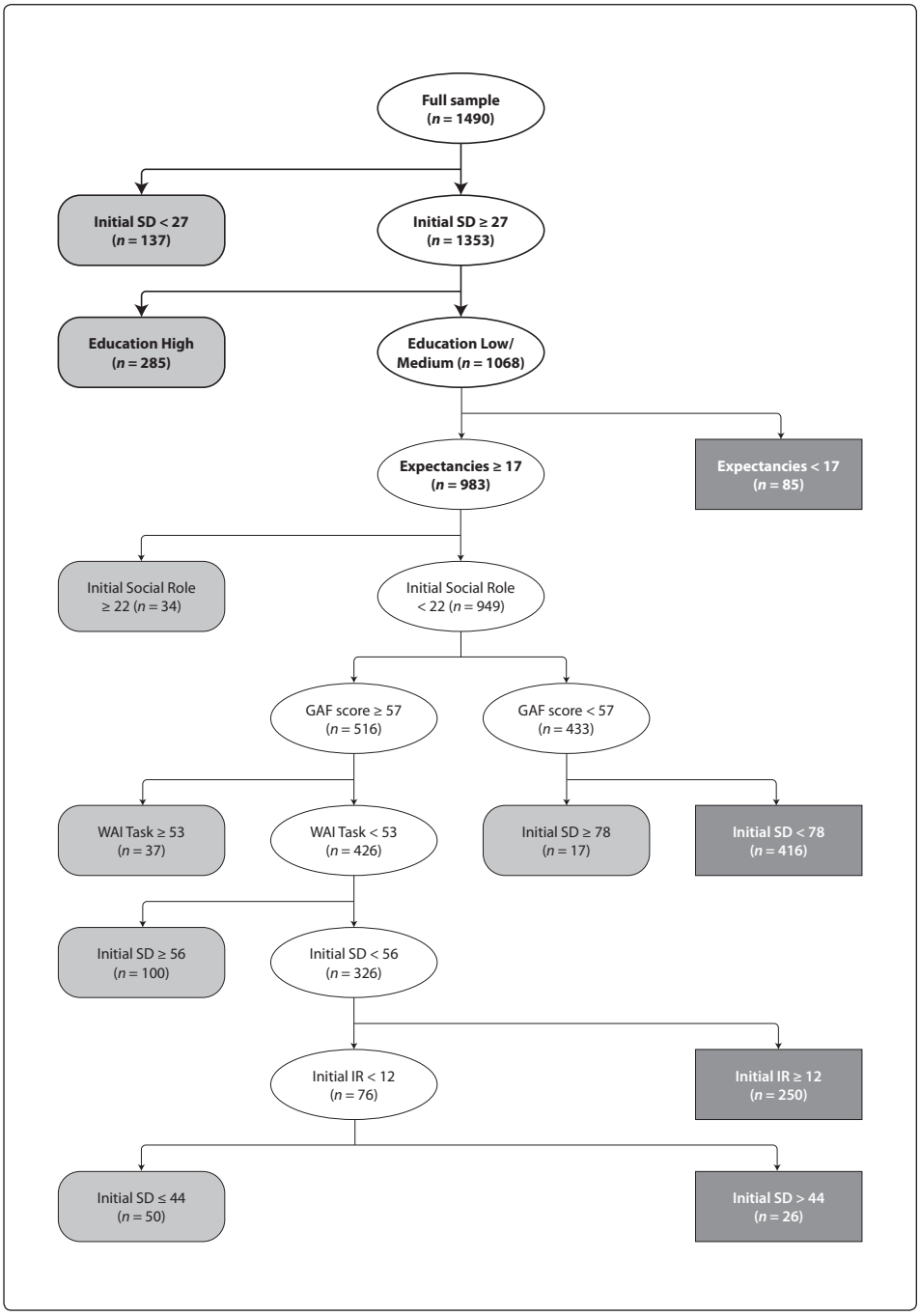
Predicting negative outcomes at end of treatment

A classification and regression tree was modelled to predict negative treatment outcomes. First, a tree that is too large was constructed. The pruning step followed by using the cross-validated error rates for different cost-complexities (see Figure 1). Even though a cross validation of 10-fold is a generally accepted as a satisfactory amount, we applied 250-fold to attain a more stabilized model. Having acquired our estimations, we selected the complexity value that was within the 1 *SE* range of the lowest error rate (three nodes). With this parameter set, the tree was pruned upward and the final model was acquired. In addition to the three nodes model, a second tree with ten nodes that also fitted the cost-complexity criterion and had a similar fit was selected (the second tree under the dotted line in Figure 1). The second model should be considered an exploratory model and was used to study more complex interactions between predictors. Figure 2 shows both models, nested within each other (the three node model in bold). As can be seen in Figure 2, the first model shows that patients who have low pre-treatment scores on the SD subscale of the OQ-45 have favorable outcomes, as do patients who have a high level of education (bachelor degree or higher). Patients with a high score on the SD subscale and a low or medium level of education are most at risk for negative treatment outcomes. It also shows that patients with low expectations of treatment outcomes – given higher initial symptom severity and lower education – have an increased risk for negative outcomes. Model 2 further explores the risk factors for negative outcomes. Patients with higher expectancies, but with more problems on social role functioning on the other hand have more favorable outcomes. From here on the model becomes more complex: Given low social role problems, high expectancies, low to medium educational level, a medium to high symptom severity, in patients with a initial GAF score below 57 (poor overall functioning), but a score on the initial SD subscale between 27 and 78 negative outcomes are more likely. Those patients who have an initial GAF above 57 and a good working alliance (WAI above 53) are predicted to have favorable outcomes, whereas a low working alliance, combined with an initial severity on the SD scale of 27-55 and high initial problems on interpersonal relationships are predicted to have negative outcomes. That is also true for patients with an initial severity on the SD scale between 44 and 55 and low initial problems on interpersonal relationships.

Predicting rate of change

Variables predicting the rate of change were investigated by using a two level multilevel analysis, with the repeated measures within patients at level 1 and differences

Figure 2 Nested CART models for negative treatment outcomes



Note: The two classification trees are nested. The smallest model is in bold, the extended model consists of all the branches that are shown. Negative outcomes are dark grey and square, positive outcomes light grey and square with rounded corners.

Table 3 The unconditional growth model and final model predicting the rate of change for the complete case analysis sample (n = 541) and the imputed data (n = 540)

		Complete case analysis	Imputed data		
	Parameter	*Estimate (SE)*	*Estimate (SE)*	*Fraction of missing information*	*Relative increase variance*
Fixed effects					
Initial status	Intercept	78.37 *** (1.03)	79.40 (0.62)	0.002	0.00
Rate of change	Intercept	-21.51 * (9.64)	-15.52 (9.42)	0.577	1.29
	Initial SD	0.80 *** (0.07)	0.38 *** (0.08)	0.631	1.61
	Initial IR	0.72 *** (0.15)	0.44 *** (0.12)	0.244	0.32
	Initial SR	0.76 *** (0.18)	0.31 * (0.14)	0.231	0.29
	GAF score	-0.25 *** (0.10)	-0.20 ** (0.07)	0.238	0.30
	WAI Task scale	-0.71 *** (0.13)			
	WAI Goal scale		-0.41 (0.12)	0.492	0.92
	Mood disorder	-3.88 * (1.83)			
	Adjustment disorder	-5.70 * (2.29)			
	Comorbidity on Axis 1		2.61 * (1.21)	0.090	0.10
	Previous treatment		3.63 ** (1.25)	0.202	0.25
	Personality disorder		4.65 * (2.22)	0.076	0.08
Variance components					
Level 1	Within-person	94.71 (3.46)	94.23 (2.41)	0.002	0.00
Level 2	Intercept	481.00 (34.69)	467.38 (21.13)	0.004	0.00
	Covariance	-550.29 (86.95)	-296.35 (42.72)	0.331	0.48
	Slope	878.10 (190.27)	467.97 (66.57)	0.339	0.50
Goodness of fit	Deviance	20900	49929		
	AIC	20926	49956		

Note: Time is modeled as the 10log of the session number. *p<0.05 *** p<0.001
Due to different sample sizes, the deviance and AIC values of the two analyses cannot be compared directly.

between patients at level 2. First, the complete case sample was analyzed, followed by the imputed data set. The final models for both analyses are presented in Table 3. As can been seen in Table 3, some predictor variables were significant in both models, but some differences could be observed as well. For pre-treatment scores on the SD, IR and SR subscales of the OQ-45 higher scores slow down the rate of change. A higher GAF score, indicating better functioning, is positively related to the rate of change. The working alliance Task scale has a positive relationship with the rate of change in the complete case sample, whereas the Goal scale is positively related to the rate of change in the imputed data set. The complete case sample emphazises the relationship with Axis I disorders, showing a more positive rate of change for patients that have mood disorders or adjustment disorders as their main diagnosis. The imputed dataset shows a stronger emphasis on the complexity of the presented problems, demonstrating that having comorbid Axis I disorders, having had previous treatment for psychological complaints and having a personality disorder as main diagnosis slow down the rate of change. Since the imputed data set most likely provides the most valid base for prediction, further analyses and interpretations were only performed for this model.

Model performance

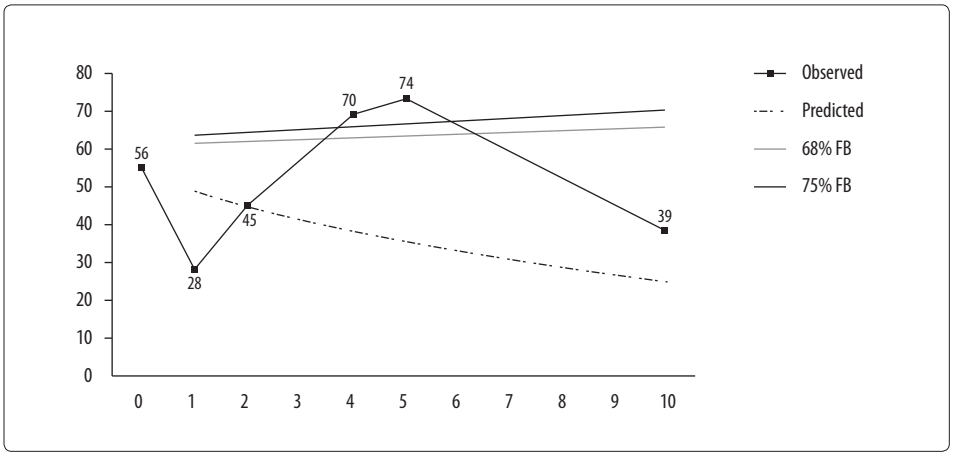
To test model performance, the sensitivity, specificity and the positive and negative predicted values were computed for both CART models (see Table 4). Model 1 had better performance in sensitivity (0.82), which means that most cases with negative outcomes are picked up, but had poor specificity (0.39), so many people that would be detected by the model as having a high risk of negative outcomes do not actually

Table 4 Model performance for the CART models and multilevel model with failure boundaries (n=1540)

	Negative outcome				Deterioration	
	CART model 1	*CART model 2*	*ML 68% FB*	*ML 75% FB*	*ML 68% FB*	*ML 75% FB*
Sensitivity	0.82 [0.80-0.84]	0.68 [0.66-0.70]	0.40 [0.38-0.42]	0.34 [0.32-0.36]	0.89 [0.87-0.91]	0.85 [0.83-0.87]
Specificity	0.39 [0.37-0.41]	0.63 [0.61-0.65]	0.78 [0.76-0.80]	0.82 [0.80-0.84]	0.75 [0.73-0.77]	0.80 [0.78-0.82]
Positive predicted value	0.55 [0.53-0.57]	0.64 [0.62-0.66]	0.62 [0.60-0.64]	0.64 [0.62-0.66]	0.25 [0.23-0.27]	0.29 [0.27-0.31]
Negative predicted value	0.70 [0.68-0.72]	0.67 [0.65-0.69]	0.58 [0.56-0.60]	0.57 [0.55-0.59]	0.99 [0.97-1.00]	0.98 [0.96-1.00]

Note: CART= Classification and Regression Tree; FB = Failure bound; ML = Multilevel analysis.
The 95% confidence intervals for the values are reported between brackets.

Figure 3 Example of the multilevel model for an individual patient



Note. FB = Failure bound

have negative outcomes. Model 2 has a better balance between sensitivity (0.68) and specificity (0.63), although both values were lower than ideally would be the case.

For the multilevel model based on the multiply imputed data sets, two confidence interval based failure bounds were computed in order to determine if deviation from the predicted treatment course (based on the multilevel model), would result in a higher risk for negative outcomes. If the failure bound was crossed the patient was considered to be at risk for negative outcomes (see Figure 3). As can be seen in Table 4, both the 68% and 75% failure bound had reasonably good specificity values (0.78 and 0.82 respectively), meaning that patients who do not cross the interval have a good chance of having positive treatment outcomes. However, specificity is poor (0.40 and 0.34), so patients at risk for negative outcomes are not picked up very well. Since we defined negative outcomes as still scoring in the clinical range and either no change or deterioration, secondary analyses were performed to test how well the multilevel model performs in identifying patients at risk for deterioration. After all, for some patients, (reliable) change may not be feasible. The multilevel model does predict deterioration well. The 68% failure bound has somewhat better sensitivity (0.89) but the 75% failure band had better specificity (0.80) while maintaining satisfying sensitivity (0.85).

Discussion

In this article, we aimed to predict risk factors for negative treatment outcomes at the end of treatment using CART and for rate of change using multilevel modeling (combined with multiple imputation). Fifty-one per cent of the patients in our

sample improved and had scores on the OQ-45 outside the clinical range at the end of treatment. In the CART analyses we found that patients with relatively low pre-treatment scores for symptom distress, and patients with high education and positive expectancies have a better chance of favorable outcomes. The extended model showed the complexity of the relation between predictors and outcome and showed how pre-treatment expectancies, social role problems and GAF scores and the working alliance at the beginning of treatment (Task subscale) interacted in different ways to predict negative outcomes at the end of treatment. The multilevel analyses showed that initial severity, the working alliance (Task or Goal subscale) and GAF score were significant predictors for the rate of change in patients. In the complete case sample, having a mood or adjustment disorder as main diagnosis had a positive relationship with the rate of change, whereas in the imputed sample previous treatment, having comorbid Axis I disorders and having a personality disorder as main diagnosis had a negative relationship with the rate of change. The model based on the multiply imputed data was considered the most reliable model, and further analyses were computed only for this model. The CART models and multilevel models differed in their sensitivity to detect negative outcomes. The first CART model had high sensitivity, but low specificity, whereas the multilevel model had high specificity and low sensitivity. The multilevel model was good at picking up deterioration, but not at identifying the no change group. The extended CART model had the best balance between sensitivity and specificity. Although sensitivity and specificity values were not as high as we are used to for instance in diagnostic tests (e.g. Gilbody, Richards, Brealey, & Hewitt, 2007), in predicting outcomes, those values are quite good, considering that there are many variables that influence the course and outcome of therapy and that most of the predictors in our study were measured pre-treatment. Our results match earlier multilevel prediction models (Lutz et al., 1999; 2001), in which the GAF score, expectancies and initial severity were found to be significant predictors, except we found some additional predictors to be significant and did not find an effect of expectancies in the multilevel analyses.

The factors that influence outcome and rate of change are in part overlapping, but not exactly the same. Expectancies and educational level were relatively strong predictors in the CART models, but not significant for predicting outcome. Having had prior treatment and having a personality disorder as main diagnosis was predictive for the rate of change, but not for outcome. Although they are highly related – people that progress fast, usually have a better chance of favorable outcomes – there are differences too. For instance, patients who have low initial severity (in the non-clinical range) in terms of symptom distress, usually have a low rate of change, but a better chance of positive outcome according to our definition. And patients who start with high initial severity and have an average progress may still have poor outcomes. We have defined

negative outcomes as being in the clinical range at the end of treatment and having experienced no (reliable) change or deterioration. This is a slightly different definition than that used by others (e.g. Lutz et al., 2006), that combined deterioration and no change as negative outcomes, including patients who functioned in the normal range at the end of treatment. We chose to define people who were functioning in the normal range as positive outcomes; since functioning for this group is comparable to people that do not have psychiatric complaints. As a result, our results differed as well. Lutz et al. (2006) found their model to be more sensitive than ours, but less specific, probably because they had a more homogeneous sample, resulting in smaller confidence intervals. The definition of negative outcome also influences the CART models, in the sense that patients who function in the normal range at the beginning of treatment usually still do so at the end of treatment. So it is no real surprise that patients with low initial severity are in the first branch of the regression tree. Patients who start just above the cut-off score for normal functioning (55) and end up just below it, and who have in fact not changed a lot, are also considered a treatment success according to our definition. However, every definition of negative outcomes has its drawbacks, including the Jacobson & Truax (1991) criteria, which are considered very conservative (Jacobson, Follette, & Revenstorf, 1984). Slight variations in the definition of negative outcomes could lead to different results in the prediction models. Sensitivity analysis could provide more insight in the extent to which these variations have impact on the results. It should be noted that the term negative outcomes has the connotation that patients could do better, but for some patients no change may be the best obtainable result. The no change group is a complicated group to start with, as it probably consists of several subgroups, including patients who have a more chronic course and are not expected to change, patients who have somewhat improved, but not enough to meet the criteria for reliable change of Jacobson & Truax (1991) and patients who come to therapy for other reasons than symptom reduction (e.g. insight or life phase problems) (Watson, 2011). This may be the reason that the multilevel model performed better in predicting deterioration than deterioration and no change combined.

In clinical practice, a combination of the CART models and multilevel model could be used to identify patients at risk for negative outcomes and to develop measures to prevent them. For instance, patients who are having high symptom distress and low or medium education may be monitored more closely during treatment, as they have an increased risk of having negative outcomes. The multilevel model can then be used to predict change. We can distinguish four groups: (1) patients who are at risk for negative outcomes and have an expected change treatment course that shows no change or deterioration; (2) patients who are at risk, but have an expected positive treatment course; (3) patients who are predicted to have positive outcomes, but have a negative

or no change expected treatment course and (4) patients who have predicted positive outcomes and also a positive expected treatment course. The first group probably has the highest risk of having actual negative outcomes. As these patients are *expected* to have a negative treatment course, the patient may not cross the failure boundary, but still have a negative treatment result. In these cases, the multilevel model might not be very effective in preventing the negative outcomes from happening, but intensive monitoring is still advisable. Other treatment options, such as seeing patients more frequently, or combining individual therapy and group therapy should be considered as well. The extended CART model could be used to identify subgroups of patients with better chances of favorable outcomes within this high risk group, and factors that also influence the rate of change, such as expectancies and the agreement on what needs to be done in therapy (Task) should be actively addressed by the therapist. For the second group, that is at risk for negative outcomes but with a favorable predicted treatment course, the multilevel model could be combined with intensive monitoring to assure that patients who go off track (and thus have a fair risk of deterioration) are identified early on in treatment. The third group, with positive predicted outcomes, but a non-positive expected treatment course probably includes patients who start with low symptom severity and are not expected to improve much. Another options is that they have favorable characteristics (e.g. low symptom severity, high education), combined with negative expectancies and a low (early) working alliance. Again, the extended CART model could provide more insight in which patients are more likely to have negative outcomes within this group. The fourth group probably has the best chances of favorable outcomes. In this group, the frequency of measurements could be decreased. We know from studies by Lambert that 30-50% of patients go off the expected track during treatment (e.g. Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004). His feedback system greatly improves outcomes for those off track patients (Lambert, 2007; Shimokawa, Lambert, & Smart, 2010), but has the consequence that patient progress has to be monitored for all patients on a session by session basis, which is a big investment of time and money. If we identify at risk patients prior to treatment, we could focus on monitoring those patients who are at risk more intensely and relax the measurement frequency for patients who have a low risk of negative outcomes and thereby improve the cost-effectiveness of the tracking system.

Using state of the art statistical techniques in analyzing our data enabled us to use all available data, and provided us with enough power to detect relevant predictor variables. The complete case sample shows that had we not used multiple imputation, results would have been different, as we would have been forced to drop some of the predictor variables, including several of the ones we were interested in (e.g.

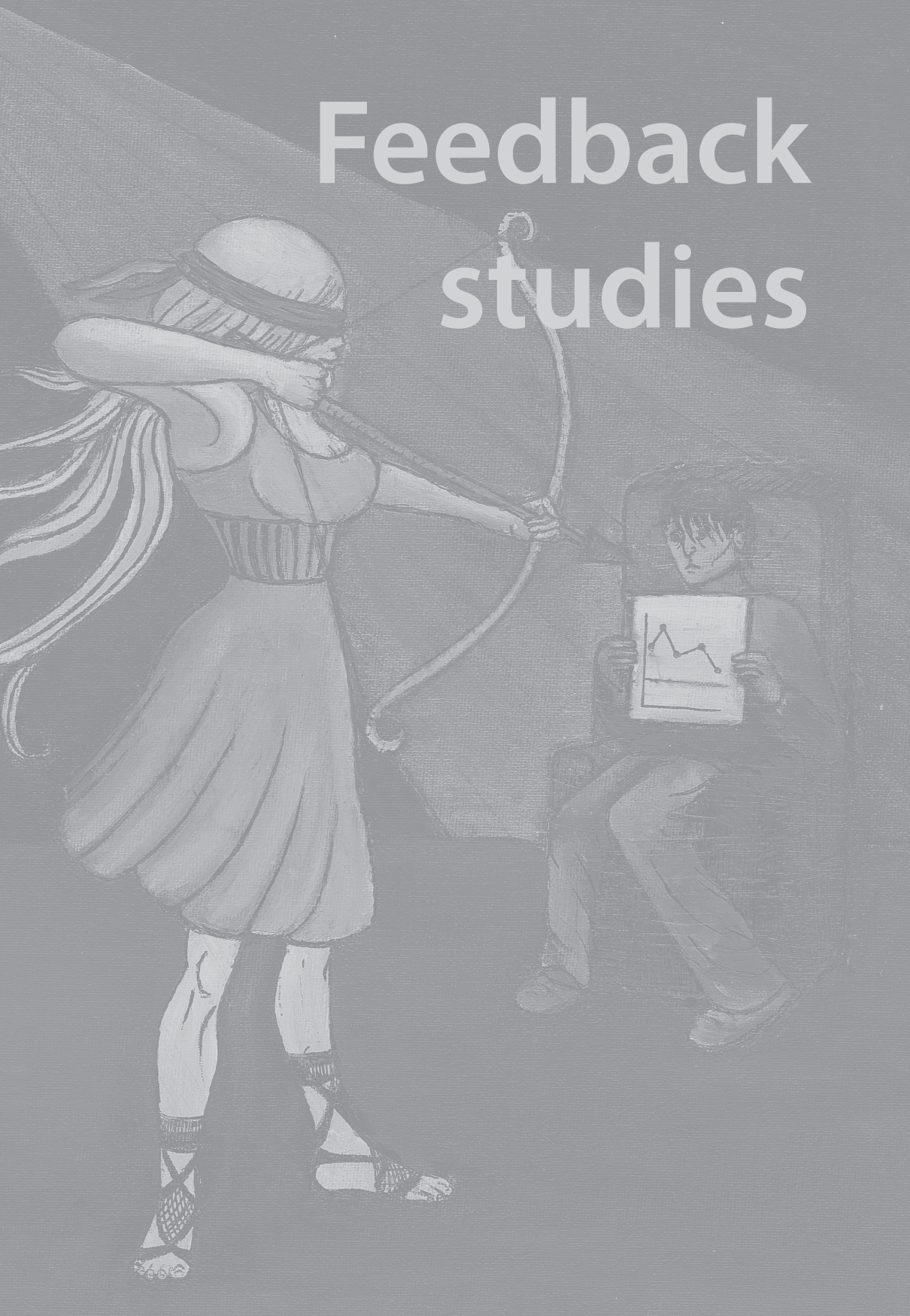
educational level, previous treatment and duration of complaints), and drop cases without complete values on the left-over predictor variables. Interesting as these methods are, like all other methods, they have their limitations as well. For instance, one of the criticisms on CART is that is not a hypothesis driven technique and therefore strongly depends on the data. Another issue is that the models usually are not very stable. That is why the models need to be cross-validated. The imputation mechanism of using surrogate splits in CART is more limited than multiple imputation, as it uses a binary value as a surrogate for the missing value (higher or lower than split value X on predictor Y), whereas multiple imputation provides a plausible value for the missing value. A major drawback is that CART cannot be performed on a multiply imputed dataset, as it would create a different tree for each imputed dataset and results could not be pooled. As a result, when outcome variables are missing, only single imputation or the last observation carried forward method can be used, both of which are not ideal. It should be noted that all methods that deal with missing data can only fix the problem to the extent that the data are Missing At Random (MAR), in other words: There should not be selective missingness that depends on unobserved variables. Although the methods applied here perform much better in Missing Not at Random (MNAR) situations than most other methods (e.g. Graham, 2009), results are still likely to be somewhat biased. Like in most naturalistic datasets, it is probable that at least part of our data was MNAR.

The main objective of this study was to develop a good prediction model that is useful in clinical practice and could be used to help prevent negative treatment outcomes. In our aim to improve outcomes for this group, an important consideration is how much outcome is to be expected. Some of the patients may simply never achieve positive outcomes. However, comparing results from randomized trials and clinical practice (Barkham, et al., 2008; Hansen, et al., 2002) suggest that in clinical practice there is still room for improvement, even though patients who participate in clinical trials may not be entirely representative of patients who are seen in everyday practice. Using prediction models in clinical practice has mainly been successful in reducing deterioration rates, but less effective in improving the outcomes for the no change group. As a field, we need to continue looking for better prediction models to help improve outcomes for this group as well and combining models that aim to predict outcomes as well as rate of change. Searching for interactions between predictive factors might be helpful in building more complex predictive models. Fortunately, statistical developments are progressing fast to help us to develop more complex prediction models suitable for analysing data from naturalistic settings.

Acknowledgements

The authors would like to thank patients and staff from GGZ Noord-Holland-Noord, GGZ Dijk en Duin and PsyQ Haaglanden that have participated in this study, in particular Patricia van Sluis, Ed Berretty and Kosse Jonker, as well as all the students that have assisted in the data collection. We would also like to thank John Ogrodniczuk for his feedback on the draft of this article.

Feedback studies



Chapter 5

Understanding the differential impact of outcome monitoring: therapist variables that moderate feedback effects in a randomized clinical trial

de Jong, K., van Sluis, P., Nugter, M.A., Heiser, W.J. & Spinhoven, P. (in press)

Psychotherapy Research.

Abstract

Providing outcome monitoring feedback to therapists seems to be a promising approach to improve outcomes in clinical practice. This study aims to examine the effect of feedback and investigate whether it is moderated by therapist characteristics. Patients (n = 413) were randomly assigned to either a feedback or a no feedback control condition. There was no significant effect of feedback in the full sample, but feedback was effective for not on track cases for therapists that used the feedback. Internal feedback propensity, self-efficacy, and commitment to use the feedback moderated the effects of feedback. The results demonstrate that feedback is not effective under all circumstances and therapist factors are important when implementing feedback into clinical practice.

Introduction

A large body of research, performed over 40 years, has demonstrated that psychotherapy can effectively improve functioning in patients (Lambert & Ogles, 2004). In randomized controlled clinical trials (RCTs) an average of 67% of patients is statistically reliably improved at the end of treatment (Hansen, Lambert, & Forman, 2002). In clinical practice, the success rates are much lower: only 35% of the patients were improved and the effect sizes of improvement were less than half the effect sizes of RCTs (Barkham, et al., 2008; Hansen & Lambert, 2003; Weisz, Donenberg, Han, & Weiss, 1995). These differences in outcomes may in part be due to selection criteria used in RCTs. However, Blais et al. (2011) found that in clinical practice the improvement rate is 57% when patients are selected that would qualify for inclusion in RCTs, which is still lower than rates found in RCTs.

Bickman (2008) considers feedback interventions a promising approach to improve clinical practice. Kluger and DeNisi (1996) define feedback interventions as actions taken by external agents to provide information regarding some aspect of one's task performance." In psychotherapy research, a common example is the monitoring of patients' progress during treatment and providing feedback to therapists on that progress. According to Bickman, clinicians need to have more systematic and reliable information about the status of their patients, in order to adjust their treatment if necessary, thus improving outcomes. In a recent review article, Carlier et al. (2010) concluded that feedback appears to have a positive impact on diagnosis and communication between patient and therapist, but effects on outcome were less clear. Meta-analyses show effects of feedback on outcome in the range of very small to very large (Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Sapyta, 2004 in Sapyta, Riemer, & Bickman, 2005; Shimokawa, Lambert, & Smart, 2010). Feedback appears to be most effective for patients that are not progressing well in therapy, the so-called not on track (NOT) cases (Lambert, et al., 2003; Sapyta, 2004 in Sapyta, et al., 2005). Carlier et al. found that feedback did not have a positive effect in 16 of the 45 (36%) mental health trials they included in their sample and Knaup et al. showed that feedback even had a negative effect in three of the twelve studies they included in their analyses.

The largest effects of feedback in mental health care have been found by the research group of Michael Lambert. They have performed five controlled studies in which therapists received feedback about a patient's improvement through the use of progress charts and warning signals about NOT cases. Results showed that NOT patients in the feedback condition had significantly more improvement than in the no-feedback control condition. The effect sizes of the various feedback conditions compared to the control conditions ranged from 0.16 to 0.70 in the full sample (Shimokawa, et al., 2010). Feedback did not have a significant effect in the on track (OT) cases (Lambert, 2007).

The feedback that the Lambert group provides to therapists is very specific. A patient's change is compared to an expected treatment course based on a statistical model and the therapist gets a warning signal when the patient deviates too much from the expected course. Most outcome monitoring systems are not as advanced. Worldwide, there are many large initiatives (e.g. Burgess, Pirkis, & Coombs, 2006; Evans, et al., 2002; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Kraus, Seligman, & Jordan, 2005; Miller, Duncan, Sorrell, & Brown, 2005; Wing, et al., 1998). Some systems have developed expected treatment recovery curves, but most do not include them in feedback on the individual patient. The existing systems vary greatly in frequency of assessment, content of the feedback and the way in which feedback is provided (Trauer, 2010). Often, feedback is not provided on a session-by-session basis, but at treatment evaluations, for instance every three months. Most feedback systems do not have signals for patients that are not progressing well in therapy. It is often assumed that all types of feedback will be effective in improving outcomes, but in fact, not much controlled research has been done on the subject (Marshall, Haywood, & Fitzpatrick, 2006).

Although it seems that feedback has potential to enhance outcomes in clinical practice, there are still many unanswered questions about how feedback works. In order to explain why outcome monitoring feedback leads to improvement in some cases and not in others, more insight is needed in the underlying processes of feedback. Characteristics of the therapists and the way in which they use feedback may play a central role in the effectiveness of feedback. After all, if therapists do not use feedback constructively, it is unlikely that it will improve outcomes. There is not much empirical knowledge on the effects that recipient characteristics have on the effectiveness of feedback (Kluger & DeNisi, 1996).

Riemer and Bickman (2011) propose the contextualized feedback intervention theory (CFIT) to explain how feedback is interpreted and used in clinical practice. CFIT focuses on the way that feedback gets attention and is accepted by therapists. When a person receives feedback a comparison is made between the content of the feedback and a goal. So if a therapist receives progress feedback, a comparison is made between the goal (recovery) and the feedback (current health status and progress so far). This comparison creates a positive or negative evaluation of the therapist's performance relative to the goal. When a discrepancy is noted, people are motivated to reduce it (Kluger and DeNisi, 1996).

This implies that behaviour change as a result of feedback will only occur if therapists attend the feedback and accept the feedback as valid (Riemer & Bickman, 2011). Feedback is more likely to be accepted if it comes from a source that has credibility and has personal relevance to the receiver (Claiborn & Goodyear, 2005). This concept is referred to as perceived validity. Another factor that seems important in acceptance is feedback

orientation (Herold & Fedor, 2003; Herold, Parsons, & Rensvold, 1996). External feedback propensity reflects the preference for externally mediated feedback as well as greater faith in such information than in what one can self-generate, whereas internal feedback propensity reflects preference for internally generated feedback as well as the tendency to reconcile differences between internal and external feedback in the direction of internally generated information. An external feedback propensity is associated with more feedback seeking behavior and better performance on novel tasks (Herold & Fedor, 2003).

Self-efficacy is another recipient characteristic that influences the feedback process. It refers to a person's beliefs concerning his or her ability to successfully perform a given task or behavior (Bandura, 1977). In case of negative feedback, people with high self-efficacy are motivated to increase their effort to reach the goal, whereas people with low self-efficacy tend to lower the goal (Kluger and DeNisi, 1996). People who have high self-efficacy also tend to consider negative feedback as more desirable than positive feedback (Claiborn and Goodyear, 2005).

Therapists' commitment to use the feedback in therapy might also be an important factor. Australian research showed that 44% of therapists thought outcome monitoring was a waste of time (Aoun, Pennebaker, & Janca, 2002) and two-thirds of the therapists were not willing to use the monitoring feedback, even if it would lead to demonstrably better outcomes (Walter, Cleary, & Rey, 1998). Riemer and Bickman (2011) state that therapists will be more committed to use the feedback if they link it to higher-level personal goals, such as being a good therapist. An a priori commitment to use feedback is expected to be highly related to actual use of the feedback.

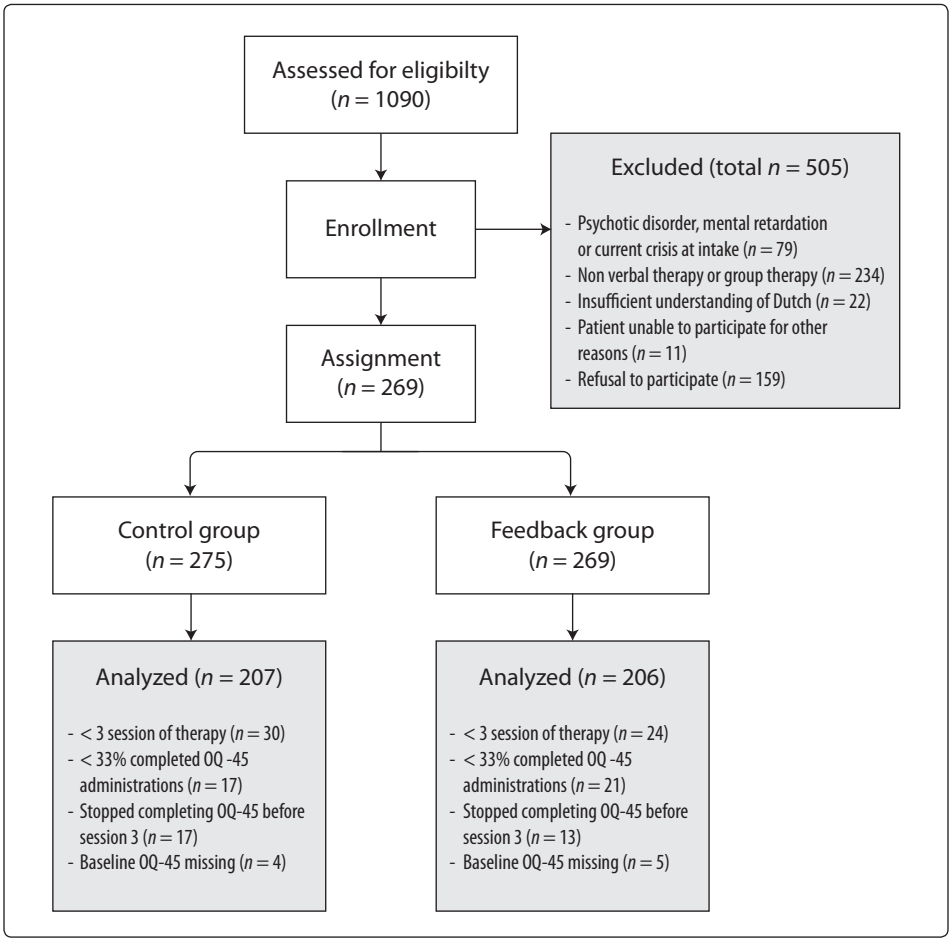
In summary, outcome monitoring has potential to improve outcomes, especially feedback with expected recovery curves and alarms for 'not on track' cases tending to result in positive effects. Most outcome monitoring systems used in clinical practice do not have these features and not much is known about the effectiveness of these systems. The effectiveness of feedback may also be related to therapist characteristics. If therapists do not accept the feedback and are not inclined to use it, feedback is not likely to be effective. In this study, we aim to research the efficacy of 'simple' (no warning signals or expected recovery curves) feedback in clinical practice compared to no feedback. Patients will be randomly assigned to a no-feedback control group or the feedback condition. We expect that patients in the feedback condition will have faster progress, compared to the no feedback control group. Although no alarms are used, therapists may be able to identify NOT cases themselves. Therefore, NOT patients will be identified post-hoc, based on reliable deterioration during the course of treatment, and it is expected that feedback will be most effective for this group. A secondary aim is to investigate whether therapist characteristics moderate the effect of feedback.

Method

Patients

During the inclusion period 1090 outpatients were screened for participation in the study in the three participating treatment departments. The treatment departments were part of two medium sized mental health care institutions in the Netherlands and typically treated a wide range of psychiatric disorders, including mood, anxiety, adjustment and personality disorders with an outpatient population. Exclusion criteria were: psychotic disorder, mental retardation, a current crisis at the time of referral, non-verbal treatment (e.g., internet therapy, pharmacological therapy, art therapy), group therapy as main treatment, re-referral within the same treatment centre within

Figure 1 Flowchart of participants



six months, and an insufficient level of understanding of Dutch. Of the remaining 703 patients, 159 declined to participate in the study.

In total, 544 patients were randomly assigned to the feedback group or control group. The first progress feedback was provided immediately before session 3; therefore, patients who had less than three sessions of therapy or stopped completing questionnaires before session 3 were excluded from analysis. Patients that had missing baseline measurements or completed less than a third of the measurements were also excluded from analysis. The flow of participants through the study is presented

Table 1 Characteristics of the patients in the feedback and control group

	Sample entered study ($n = 544$)				Sample in analysis ($n = 413$)			
	Not in analysis		In analysis		Control		Feedback	
	n	%	*n*	%	*n*	%	*n*	%
Sex								
- Female	131	75 (57%)	413	252 (61%)	207	124 (60%)	206	128 (62%)
Age	131	M = 37.0 SD = 12.3	413	M = 36.8 SD = 11.9	207	M = 36.9 SD = 11.8	206	M = 36.7 SD = 12.1
Marital status								
- Single	128	58 (45%)	410	178 (43%)	206	94 (46%)	204	84 (41%)*
- Living together		10 (8%)		32 (8%)		14 (7%)		18 (9%)
- Married		39 (31%)		136 (33%)		62 (30%)		74 (36%)
- Divorced		21 (16%)		56 (14%)		35 (17%)		21 (10%)
- Widowed		0 (0%)		8 (2%)		1 (0.4%)		7 (3%)
Education								
- Low	106	39 (37%)	380	94 (25%)*	193	48 (25%)	187	46 (25%)
- Medium		57 (54%)		226 (60%)		113 (59%)		113 (60%)
- High		10 (9%)		60 (16%)		32 (17%)		28 (15%)
Main DSM-IV disorder								
- Mood	131	35 (27%)	413	96 (23%)	207	47 (23%)	206	49 (24%)
- Anxiety		17 (13%)		79 (19%)		34 (16%)		45 (22%)
- Adjustment		32 (24%)		92 (22%)		50 (24%)		42 (20%)
- Personality		9 (7%)		31 (8%)		16 (8%)		15 (7%)
- Eating		0 (0%)		10 (2%)		8 (4%)		7 (3%)
- Usually first diagnosed in childhood		7 (5%)		13 (3%)		6 (3%)		7 (3%)
- Substance related		2 (2%)		12 (3%)		7 (3%)		5 (2%)
- Somatoform disorder		2 (2%)		13 (2%)		5 (2%)		8 (4%)
- Impulse control		2 (2%)		10 (2%)		7 (3%)		3 (2%)
- Other		24 (18%)		46 (11%)		24(12%)		22 (11%)
- No Axis I or II disorder		1 (1%)		6 (2%)		3 (1%)		1 (2%)
Comorbidity								
- Multiple Axis I disorders	131	52 (40%)	413	154 (37%)	207	76 (37%)	206	78 (38%)
- Comorbid Axis I and II disorders		26 (20%)		97 (24%)		50 (24%)		47 (23%)
OQ-45 intake score	127	M=72.9 SD= 23.4	413	M= 76.7 SD= 22.1	207	M=76.7 SD= 22.0	206	M= 76.7 SD= 22.3
Prior treatment								
- Yes	127	75 (59%)	410	237 (58%)	206	124 (60%)	204	113 (55%)

Note
* $p < .05$

in Figure 1. The 413 patients that were included in the analysis included 252 females (62%), aged 18-64 years (*M* = 36.8; *SD* = 11.9). Patient characteristics are reported in Table 1.

Therapists

There were 57 therapists who participated in this study, 21 males (37%) and 36 females (63%), aged from 26 to 60 years with a mean age of 45.3 years (*SD* = 9.7). Therapists were psychologists (49%), psychiatric nurses (39%), social workers (7%) or other mental health care professionals (5%). The therapists had 0 to 35 years of experience after getting licensed, with a Mean of 13.8 years (*SD* = 11.3). Therapists had between 1 and 23 clients in the study (*M* = 7.3; *SD* = 5.6). Therapies provided included cognitive behavioral therapy, interpersonal therapy, brief solution focused therapy and counseling. Most therapies were integrative and did not represent a single therapy orientation.

Instruments and manipulation

Outcome Questionnaire-45 item version (OQ-45)

The Outcome Questionnaire-45 (OQ-45) was used to measure patient progress during treatment. The OQ-45 (Lambert, et al., 2004) is a self-report instrument and has 45 items, 9 of which are reversed, asking how the respondent has felt over the last week on a 5 point rating scale, ranging from 0 (never) to 4 (almost always). The OQ-45 consists of three subscales: Symptom Distress, Interpersonal Relations, and Social Role. The Symptom Distress domain consists of 25 items relating to psychological symptoms that are common in highly prevalent mental disorders. The Interpersonal Relations domain consists of 9 items that assess functioning in interpersonal relationships, and the Social Role domain consists of 11 items that assess functioning in social roles, such as work and school. The cutoff score for normal functioning is 55 for the Dutch OQ-45 and the reliable change index is 14. The internal consistency for the Dutch version of the OQ-45 is between 0.92 and 0.96 for the Total Score in university, community, patients and community and patients combined samples. For the subscales, the internal consistency is 0.90-0.95 for the Symptom Distress scale, 0.74-0.84 for the Interpersonal Relations subscale and 0.53-0.72 for the Social Role subscale (De Jong, Nugter, Lambert, & Burlingame, 2009).

Demographic questionnaire

The demographic questionnaire is a 19-item self-constructed questionnaire that assesses the demographic characteristics of the patient. It asks for the patient's date

of birth, gender, postal area code, nationality, country of birth, country of birth of the patient's parents, marital status, living and working situation, educational level, prior treatment, pretreatment use of medication, the main complaint, and the duration of the main complaint.

Feedback User Questionnaire

This questionnaire consisted of the Internal and External Feedback Propensity Scales and an adaptation of the CFIT User Survey.

The *Internal and External Feedback Propensity Scales* (IEFPS; Herold, et al., 1996) are used to measure feedback propensity. The instrument consists of two subscales that measure internal and external feedback propensity. Each subscale consists of six items that are answered on a five-point rating scale that varies from strongly disagree to strongly agree. An item from the External Feedback Propensity Scale is 'It is very important to me to know what people think of my work.' A sample item from the Internal Feedback Propensity Scale is 'How other people view my work is not as important as how I view my own work.' The reliability of the IEFPS was 0.71 for the external feedback propensity scale and 0.73 for the internal feedback propensity scale (Herold, Parsons, & Fedor, 1997). In our sample, the internal feedback propensity scale had a Cronbach's α of 0.71 and the external feedback propensity scale had an α of 0.62.

An adaptation of the *CFIT User Survey*, designed by the Center for Evaluation and Program Improvement of Vanderbilt University, was used to measure commitment to use the feedback, self-efficacy and perceived validity of the feedback. The items are scored on various five-point rating scales. The commitment to use the feedback was measured with a scale based on the *Goal Commitment Scale* (Hollenbeck & Klein, 1987) that consists of seven items. A sample item is 'It is hard to take the idea of using these measures in my clinical practice seriously.' The self-efficacy scale consists of eight items. A sample item from that scale is 'To what extent do you feel confident in your ability to know what to do if a client is not progressing in treatment.' Perceived validity of the feedback was measured by a six-item scale. A sample item is: 'I think that feedback based on the OQ-45 will be helpful for my counseling.' The internal consistency in the current sample was 0.90 for the commitment scale, 0.88 for the perceived validity scale, and 0.82 for the self-efficacy scale.

Use of feedback

In the original study design the use of feedback was asked per patient, but due to problems in the software, this questionnaire was not administered. Therefore, the use of feedback by the therapist was assessed post hoc by asking by e-mail if the therapists had used the feedback with their patients (yes/no) and in what way (open question). Therapists that had used the feedback usually did so in multiple ways, including

discussing the feedback with patients, giving homework assignments, and using the feedback to end the therapy when sufficient progress was made.

Feedback intervention

In the feedback condition, the therapist received e-mails that contain a progress report after sessions 1, 3, 5, and subsequently every fifth session. The patient's progress on the OQ-45 Total Score was shown in a graph. A table showed the patient's baseline score, the last available measurement, the change in scores on the OQ-45, and the clinical and reliable change status (see Appendix B). Patients were classified as deteriorated if their OQ-45 score increased 14 points or more compared to baseline on the OQ-45 Total Score and classified as reliably improved if their Total Score had decreased 14 points or more. If patients improved reliably and crossed the cut-off point for normal functioning (55), they were classified as clinically significantly changed. Patients that did not meet these criteria were considered unchanged. Positive changes were shown in green, negative changes in red. A second graph and table displayed the subscale scores. The critical items on the OQ-45 that alert the therapist to suicidal thoughts, aggression and drugs and alcohol use, were presented if patients answered them with a score of 1 (seldom) or higher. Prior to the study, all therapists were given training on how to interpret the feedback, but were given no specific guidelines to identify 'not on track' patients, consistent with the concept of simple feedback. They also received an instruction card that explained all elements of the feedback report.

Procedure

Patients were screened for eligibility after intake and contacted by phone if they did not meet the exclusion criteria. If patients agreed to participate in the study, they signed an informed consent form and received explanation using the on-site test computer. Patients completed the OQ-45 on the computer prior to each of the first five sessions of therapy, and subsequently every fifth session for a maximum period of one year. At the first session, patients also completed the demographic questionnaire. If patients were assigned to therapies that were excluded from the study, data collection was stopped and measurements up until that point were used in analysis. Therapists completed the Feedback Questionnaire prior to the study. The use of feedback questionnaire was e-mailed to the therapists after completion of the study.

Table 2 Therapist variables (n = 57)

	% missing	*Mean*	*SD*
Self-efficacy therapist role	21%	27.2	2.6
Internal feedback propensity	23%	19.6	3.4
External feedback propensity	23%	19.7	2.4
Perceived validity	18%	21.2	3.4
Commitment to use feedback	16%	23.9	3.9

Analysis

Missing data

Missing data on the therapist level were imputed using the Multiple Imputation procedure in PASW Statistics 18.0 (SPSS, 2009). The multiple imputation procedure is based on the Multiple Imputation by Chained Equations (MICE; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) algorithm. Since multiple imputation is not supported for three-level models yet, single imputation of the missing values was selected by setting the imputation number as one. Only missing values on the therapist variables were imputed. Table 2 reports the percentage of missing data and the mean score and standard deviation in the original data.

Definition of not on track

Patients that were NOT were identified post hoc. Patients were considered NOT if they deteriorated, defined by an increase in the Total Score at least as large as the reliable change index (14 points) compared to the baseline measurement at any point in their treatment. This criterion was chosen since it was mentioned in the feedback report if the patient deteriorated. A total of 67 patients (16%) were NOT according to this definition.

Main hypotheses

Baseline differences in demographic and clinical characteristics between the treatment conditions were tested with chi-square tests and independent sample t-tests using PASW Statistics 18 (SPSS, 2009). The main hypotheses were tested by two three-level multilevel models, using the PROC MIXED procedure in SAS (SAS Institute Inc, SAS 9.2. Cary, NC, USA, 2008). As time variable the 10log of the session number was used to allow for a linear model (also see Lutz, Martinovich, & Howard, 1999). Maximum likelihood estimation was used to estimate the model parameters, using an unstructured variance structure. A random intercept, random slope model (on both patient and therapist level) was used to test the main hypothesis on the effect

of feedback. First, an unconditional growth model was postulated, and then the main effect of feedback was added to the model, followed by the interaction with being NOT and use of feedback by the therapist. Redundant factors were eliminated from the model, in order to obtain a parsimonious model. To test for therapist effects, a model with a fixed slope on level two and a random slope at level 3 was used (with random intercepts). A backwards procedure was applied, starting with a full model including all relevant level 2 and 3 predictor variables and their interactions and eliminating non-significant factors (using the Wald test for fixed effects) one by one until a parsimonious model was reached, that was not significantly worse than the full model (compared with the deviance test).

To predict which therapist characteristics would predict use of feedback, a logistic regression analysis was performed in PASW statistics (SPSS, 2009), using a backwards procedure.

Results

Baseline differences between groups

Baseline differences on gender, age, marital status, education, DSM-IV disorder, OQ-45 intake score and prior treatment between the patients that were included and excluded in analysis were tested. The groups did not differ on most variables (see Table 1), except for educational level. Patients that were excluded from analysis were more likely to have a low education than patients that were included, $\chi^2(6) = 11.4$, $p = .039$. For patients that were in analysis, baseline differences between the control group and feedback group were tested. The two groups only differed on marital status: patients in the control group were more likely to be widowed and less likely to be divorced than the feedback group, $\chi^2(2) = 7.2$, $p = .027$.

Effect of feedback

There was no significant effect of feedback on the rate of change (see Model A, Table 3). The interaction between feedback and being NOT was also not significant. Adding the interaction with use of the feedback to the model revealed that for therapists that used the feedback (46%, representing 57% of the patients), there was a significant positive effect of feedback in NOT cases (see Model B, Table 3), although the effect was not large enough to counterbalance the negative change trajectory that NOT patients typically have.

Table 3 3-level models on the effect of feedback and moderating therapists factors

		Model A	Model B	Model C
	Parameter	*Estimate (SE)*	*Estimate (SE)*	*Estimate (SE)*
Fixed effects				
Initial status, β_{0j}	Intercept	77.47 *** (1.11)	77.49 *** (1.11)	77.30 *** (1.11)
Rate of change, β_{1j}	Time	-16.83 *** (1.63)	-17.57 *** (1.64)	-5.11 (9.69)
	Feedback	0.59 (2.19)		
	Not on Track	27.37 *** (3.43)	27.84 *** (2.92)	19.55 *** (1.80)
	Feedback * NOT	-5.87 (4.83)		
	Use *Feedback		3.06 (2.30)	
	Use * Feedback * NOT		-10.77 * (5.15)	
	Internal feedback propensity			0.58 * (0.27)
	Commitment to use feedback			-0.84 ** (0.29)
	Self-efficacy* Feedback			-1.24 *** (0.31)
	Commitment * Feedback			1.33 *** (0.34)
Variance components				
Level 1	Within-person	92.44	92.50	112.91
Level 2	Intercept	440.74	440.16	417.69
	Slope	175.78	172.67	
	Covariance	-56.27	-56.87	
Level 3	Intercept	0.16	0.11	2.87
	Slope	5.02	5.47	21.51
	Covariance	-6.46	-6.48	-16.19

Note: Time is modeled as the 10log of the session number. *$p<0.05$, **$p<0.01$, *** $p<0.001$
NOT = Not On Track
Negative values in the fixed part of the models correspond with a faster decrease of dysfunctioning over time.
Use, NOT and Feedback were coded as dummy variables (0 = no, 1= yes).

Table 4 Logistic regression analysis predicting use of feedback

				95% confidence interval	
	B	*SE*	Odds ratio	Lower bound	Upper bound
Constant	-5.23	2.21	0.01		
Commitment to use feedback	0.25*	0.10	1.28	1.06	1.54
Female	-6.61**	2.48	4.01	1.10	14.45

Note: $R^2 = .18$ (Cox & Snell), .23 (Nagelkerke). Model $\chi 2$ (2) = 10.78, $p = .005$. * $p<.05$, ** $p<.01$

Therapist characteristics

The effects of therapist characteristics on the rate of change are presented in Model C, Table 3. The correlation between the therapist characteristics was below $r = 0.45$ for all pairs, except for perceived validity and commitment to use feedback ($r = 0.70$). Having an internal feedback propensity had a negative effect on the rate of change, regardless of whether the therapist received feedback. A higher commitment to use the feedback had a general positive effect on the rate of change, but there was also a significant interaction between commitment and feedback in a negative direction, indicating that when therapists actually received feedback, having a higher commitment was predictive of a slower rate of change in their patients. Finally, there was a positive effect of self-efficacy in the feedback condition. Patients of therapists with higher self-efficacy expectations who received feedback had a higher rate of change than patients of therapists with lower self-efficacy expectations that did not receive feedback. Against our expectations, there was no significant effect of external feedback propensity and perceived validity.

Predicting use of feedback

Since use of feedback by the therapist significantly interacted with the effect of feedback in NOT cases, we were interested in which variables predicted use of the feedback by the therapists. Table 4 shows that a higher commitment to use the feedback increased the odds that therapists would use the feedback. Being a woman also increased the odds of using the feedback: female therapists were four times more likely to use the feedback than male therapists. No significant effects were found for the type of therapist, years of experience of the therapist, self-efficacy, internal and external feedback propensity, and perceived validity.

Discussion

This study aimed to assess the effect of monitoring outcomes and providing feedback to therapists on the rate of change in patients. Contrary to our expectations, for the full sample no beneficial effect of feedback was found and there was no significant interaction between feedback and patients being NOT. However, in NOT cases a positive significant effect was found when therapists indicated that they used the feedback. Therapist variables moderated the effectiveness of feedback. Therapists with a high internal feedback propensity, who are more likely to trust their own opinion than feedback from external sources, had patients with a slower rate of change than therapist with a low internal feedback propensity, whereas therapists who were more committed to use the feedback at the beginning of the study had patients who progressed faster. These two results suggest that therapists with an open attitude towards getting feedback reach faster progress with their patients. Strangely though, when therapists with a high commitment to use the feedback actually received feedback, this slowed down the rate of change in their patients. There also was a positive effect of self-efficacy. Patients in the feedback condition whose therapist had higher self-efficacy progressed quicker in therapy than patients whose therapist had lower self-efficacy or patients whose therapist did not receive feedback. No effect was found for external feedback propensity and perceived validity. Therapists were more likely to use the feedback if they were more committed to use the feedback at the start of the study and if they were female.

Our results demonstrate that feedback may not be effective under all circumstances for all therapists. This is in line with a recent study by Lambert's group, in which they used complex feedback in a hospital-based outpatient clinic and found much lower effects of feedback than in previous studies. Further analysis showed that feedback was only effective for half of the therapists (Simon, Harris, & Lambert, 2011). The therapists that were participating in that study had very heavy caseloads and seemed demoralized by organizational changes. Riemer and Bickman (2011) stress that organizational factors, such as a high administrative workload, can become barriers for therapists to use feedback. In a recent survey that included many of the therapists participating in the current study, therapists indicated lack of time and other tasks that were competing for their attention as important barriers to use the feedback (De Jong, in press). Londen, Smither and Adsit (1997, in Riemer & Bickman, 2011) state that if there is no accountability for using feedback, it will have little impact. Accountability should be handled with care though, as it can also provoke defensive reactions in therapists (Riemer & Bickman, 2011). Although the managers of the participating departments were actively involved in the study, it was still complicated to hold therapists accountable for using the feedback within the context of a research project.

Managers were not allowed to view therapists' progress curves, in order to prevent defensive reactions.

Some of the choices we made in designing the study could have influenced results. The chosen frequency of measurements and feedback reports was not on a session-by-session basis, as Lambert does, which may reduce the chance to signal a patient being not on track and as a result might reduce the effect of the feedback. We encountered a relatively low rate of signal cases (16%) in this study. Another issue is that patients completed questionnaires up until one year or until they were referred to treatments that were excluded from this study. As a result, a portion of the patients in our dataset stopped with the study before the end of treatment, which may have reduced the feedback effect. However, one would still expect an effect of feedback for the sessions on which feedback was provided. A third factor that may have influenced results in a negative way was several problems we occasionally encountered with the feedback software. One of the problems was that the questionnaires we had planned to administer on use of feedback did not work, which forced us to measure therapists' use of feedback post hoc. This may have several disadvantages, as therapists may not always remember accurately if they used the feedback or not. Also, demand characteristic may play a role. Therapists are aware that they should have used the feedback and may be less likely to report that they did not. However, considering that half of the therapists indicated that they had not used the feedback, we believe that the effect of sociably desirable answers was limited in our study. A final issue is that the sample may be selective to a certain degree. It seemed that patients not included in analysis were more likely to have low education, which is consistent with results from dropout studies (Clarkin & Levy, 2004). Therefore, our results may be less representative for lower educated patients.

What implications do these results have for clinical practice? It seems important to realize that not all types of feedback may be equally effective. People often refer to Lambert to justify implementing feedback, but take out elements of his feedback system that may be particularly effective. Warning signals might be effective in getting therapists' attention to look at the feedback and the statistical model underlying the expected recovery curves may cause the therapists to perceive the feedback as more valid. Another implication is that therapists' commitment to use the feedback seems to influence the feedback's effectiveness. It is especially important to pay attention to commitment to use the feedback, which predicts both rate of change and likelihood to use feedback, when implementing feedback into clinical practice. Unwillingness to use feedback may be due to uneasiness regarding receiving feedback on one's performance. After professional training and licensing, therapists no longer receive structured feedback on their performance. Not using the feedback might be a way to cope with the anxiety of not being a good therapist. That is consistent with our finding

that therapists with higher self-efficacy were able to use the feedback to their benefit, although self-efficacy was not a significant predictor of actually using the feedback. An alternative interpretation is that not all therapists might be interested in enhancing their therapeutic skills. In any case, it is important to pay attention to the role of the therapists and their use of outcome monitoring feedback when aiming to use outcome monitoring as a tool to improve clinical outcomes. Under pressure of third parties, such as health insurance companies, many organizations just start measuring and do not pay sufficient attention to how feedback works and how therapists can effectively use it (De Jong, 2012).

This is the first study that has measured therapist factors in the context of a feedback intervention. Our results demonstrated that therapist characteristics are relevant and more research in this area is needed. Therapist characteristics that might be interesting to study include attribution style, locus of control, personality traits of the therapists and emotional stability. Therapist characteristics might be manipulated by training therapists in specific feedback-related skills. In addition, it would be important to get more insight in the dynamics of how feedback works and for whom. Perhaps some groups of therapists or patients perform worse when they are provided with feedback, but so far, we do not know if this is the case. Finally, for the further development of feedback, it is crucial that the premises of feedback theory are tested in a clinical context, since most of these theories originate from social and organizational psychology. Feedback effects are considered context specific and currently the contextualized feedback intervention theory (CFIT; Riemer & Bickman, 2011) is the only available theory that focuses on clinical practice. CFIT is complex and for the largest part untested, therefore alternative theoretical models could be explored as a basis to generate new hypotheses about how feedback works in clinical practice.

Acknowledgements

The authors would like to thank patients and staff from GGZ Noord-Holland-Noord and GGZ Dijk en Duin that have participated in this study, as well as the students that have assisted in the data collection, in particular Lisanne Boom. Thanks are also due to Andrew McAleavey for his comments on the draft of this chapter.

Chapter 9

The effect of outcome monitoring feedback to clinicians and patients in outpatient mental health: randomized controlled trial

De Jong, K., Timman, R.,
Hakkaart-van Roijen, L., Vermeulen, P.,
Kooiman, K., Passchier, J., van Busschbach, J.

Manuscript submitted for publication.

Abstract

Background. Outcome monitoring has become popular, but meta-analyses show mixed results. Feedback to so called 'not on track' (NOT) cases and to both patient and therapist seems most effective.

Aims. This study aimed to evaluate the effect of outcome monitoring feedback to therapists and patients on outcome.

Method. Patients (n = 474) were randomly assigned to three conditions: feedback to therapist (FbT), feedback to therapist and patient (FbTP) and no feedback (NFb).

Results. In the full sample, no significant effect of the FbT condition was found. FbT did result in less negative change pattern for NOT cases in short-term therapies (<35 weeks). FbTP was preventive of negative change for NOT cases in short-term therapies (d = 1.28 after 35 weeks) and had a small positive effect on the rate of change in long-term therapies (≥ 35 weeks).

Conclusions. FbTP results in faster progress, especially for NOT cases. FbT was only effective in short-term therapies.

Introduction

Providing outcome monitoring feedback to clinicians and patients has become an increasingly popular method to improve outcomes and has been adopted by many mental health care providers all over the world (e.g. Evans et al., 2002; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Kraus, Seligman, & Jordan, 2005; Miller, Duncan, Sorrell, & Brown, 2005). Research has shown that measuring outcomes and providing feedback as part of routine practice appears to have a positive impact on the accuracy of diagnosis (Carlier et al., 2010; Marshall, Haywood, & Fitzpatrick, 2006) and on communication between patient and clinician (Carlier et al., 2010), but the impact on patient outcome and treatment duration is less consistent. Meta-analyses on the effect of feedback on outcome show mixed effects.

A meta-analysis by Knaup et al. (Knaup, Koesters, Schoefer, Becker, & Puschner, 2009) concluded that health status feedback has a small positive effect on outcome in short-term treatments (d = 0.10), but not in longer term treatments (d = -0.06). A problem in the meta-analysis by Knaup et al. is that compared short-term and long-term effects of feedback, but the studies in these two groups differed substantially in patient population and frequency of the feedback. The long-term therapy group consisted mainly of studies conducted in severe mental disorders and infrequent feedback (once or twice), whereas the short-term group consisted of studies in mood and anxiety disorders and personal concerns, and most studies used weekly feedback. Lambert et al. (Lambert, Whipple, Hawkins, Vermeersch, Nielsen & Smart, 2003) and Shimokawa et al. and (Shimokawa, Lambert, & Smart, 2010) found much larger effects of feedback on outcome, ranging between 0.28 and 0.70, but their meta- and mega-analysis included only studies from their own research group and were mainly conducted in the university counselling center. Feedback seems mainly effective for patients who are not doing well in therapy, the so called 'not on track' (NOT) cases (Carlier et al., 2010; Lambert et al., 2003). Not on track cases are typically identified as being those individuals who fall below a cut-off indicating an expected treatment response. There are also some indications that feedback is more effective when both the therapist and the patient receive feedback (Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004), but in other studies there was no significant additional effect (Harmon, Hawkins, Lambert, Slade, & Whipple, 2005; Slade, Lambert, Harmon, Smart, & Bailey, 2008).

The current study investigates the effect of feedback in a sample of outpatients treated in mental health care institutions or private practices. Patients completed session-by-session questionnaires in a web-based application. The main research question was whether feedback improves outcomes and whether feedback to patients and therapists would be more effective than feedback to therapists alone. There were

three conditions: feedback to therapists, feedback to patients and therapists and a no feedback control group. The feedback was expected to be mainly effective for NOT cases. We were interested in the effects of feedback in both short-term and long-term therapies. Short-term and long-term therapies were defined post-hoc by splitting on the median of treatment duration (35 weeks).

Method

Subjects

Patients

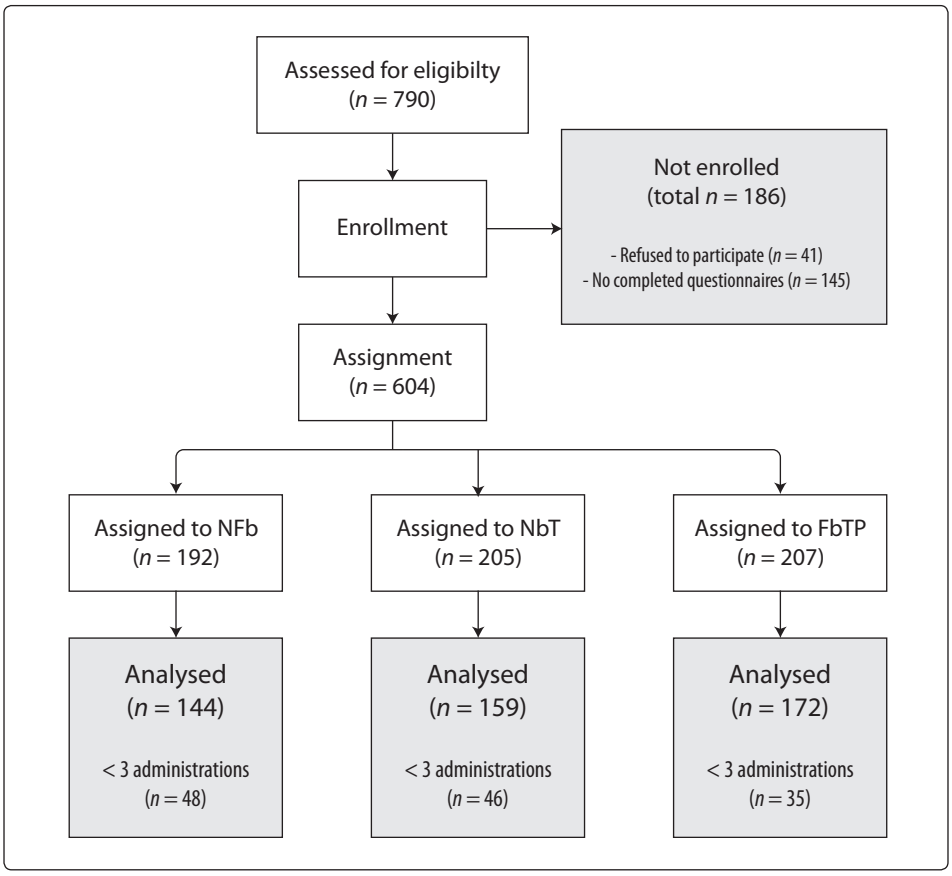
Data were collected in a web-based monitoring application in the period of July 1, 2006 to June 31, 2011. Participants were recruited in private psychotherapy practices and outpatient mental health institutes. Inclusion criteria were an age of 17 years or older and sufficient understanding of the Dutch language to answer questionnaires without assistance. Therapists asked their patients to participate in the study at intake. After agreeing to participate, subjects were randomly assigned to one of three conditions: Feedback to therapists (FbT), feedback to both therapists and patients (FbTP) or a control group without feedback (NFb). A randomized block design, with blocks within therapists, was applied. The study was approved of by the Medical Ethical Committee of the Erasmus University Medical Center Rotterdam, as well as by the cooperating institutes. All participants signed an informed consent form.

Participants with less than three OQ-45 administrations were excluded from analyses, because two administrations are the absolute minimum to present feedback with a gain or decrease that can have an effect on treatment outcome at session three or later. Figure 1 shows the flow of participants through each stage of the trial.

Therapists

A total number of 110 therapists participated in the study on a voluntary basis. In the analysed sample, therapist had between 1 and 34 patients participating in the study, with a mean of 4.3 patients per therapist ($SD = 6.4$). Approximately half of the therapists worked in private practice and most therapists were originally trained as psychologists or psychiatrists. Years of experience after training varied from 0 to 36 years, with a mean of 16.9 ($SD = 9.5$) years. Therapists of all major therapy orientations participated in the study, although cognitive behavioural therapy, client-centred therapy and psychodynamic therapy were most frequent. Table 1 shows the characteristics of the participating therapists.

Figure 1 Flowchart of participants



Instruments

Outcome Questionnaire-45 item version (OQ-45)

The Dutch version of the Outcome Questionnaire-45 item version (OQ-45) was used to measure patient progress during treatment. The OQ-45 (Lambert, et al., 2004) is a self-report instrument and has 45 items, 9 of which are reversed, asking how the respondent has felt over the last week on a 5 point rating scale, ranging from 0 (never) to 4 (almost always). Higher scores reflect a higher level of dysfunctioning. The OQ-45 consists of three subscales that are aimed at assessing different domains of client functioning: Symptom Distress, Interpersonal Relations and Social Role. The internal consistency for the Total score of the Dutch OQ-45 ranges between 0.92 and 0.96 in university, community, patients and community and patients combined samples. For the subscales the consistency is 0.90-0.95 for the Symptom Distress scale, 0.74-0.84 for the Interpersonal Relations subscale and 0.53-0.72 for the Social Role subscale (De Jong, Nugter, Lambert, & Burlingame, 2009).

Table 1 Therapists characteristics

	n	
Female	110	64%
Age	98	$M = 47.8$, $SD = 10.9$
Member of therapists association(s)	101	30 %
Private practice	108	49 %
Number of therapy hours/week	98	$M = 26.6$, $SD = 11.1$
Years of practice	87	$M = 16.9$, $SD = 9.5$
Education	103	
Psychology		63 %
Medicine		14 %
Educational sciences		8 %
Other university degree		16 %
Other education		15 %
Most important professional orientation	98	
Cognitive behavioural		27%
Client centered		24%
Psychodynamic		14%
Integrative		8 %
Eclectic		7 %
Systemic		7 %
Other		9 %

Patient characteristics

Patients completed a basic background questionnaire after entering the study. The questionnaire consisted of six items on gender, age and e-mail address of the patient, the name of the therapist and the frequency of visits to the therapist.

Clinical diagnosis

A psychiatric classification according to the Diagnostic and Statistic Manual of Mental Disorders IV on all five axes was provided by the therapist in the online system.

Procedure

The background questionnaire was administered prior to the first therapy session. Before each therapy session, though not more than once a week, the patient filled out the OQ-45 online, through a secure internet connection. Patients were provided with an individual login and password and were able to log in from any location,

although most completed their questionnaires in the waiting room of the therapist. Feedback was generated immediately for use in the therapy session. Therapists could access the feedback either through e-mail or by logging into the therapist portal of the online feedback system. Therapists and patients were free to discuss the feedback messages or not. Feedback consisted of a progress graph and a message tailored to the status of patients. The graph represented the total OQ-45 score and the subscale scores at the various therapy sessions. A horizontal red line indicated the cut-off score (i.e. 58) between the normal and clinical population. Messages to therapists included suggestions on the level of complaints and continuation of the therapy, for instance "Your patient shows a high level of complaints, but feels better than at the start of treatment. Your patient has a good chance to benefit from further treatment." In the patient feedback, patients received the same feedback as the therapist, except the feedback messages used language that was directed towards the patient.

Statistical analysis

Not On Track cases were defined by a deterioration of at least the reliable change index (14 points) compared to baseline at least twice in the course of the therapy, to ensure that a patient was really not on track and not just had one negative outlier once. Therapies were divided into short-term and long-term therapies post hoc using the median of the treatment duration (35 weeks), thus creating two groups of similar size.

Data were analyzed with multilevel modelling, using the MIXED procedure in SAS (SAS Institute Inc, SAS 9.2. Cary, NC, USA. 2008). Initially, three levels were postulated: therapists as upper level, patients as second level, and time-points as lowest level. Bias caused by very long therapies was avoided by deletion of data after 2 years of therapy (104 weeks). The deviance statistic was used for testing the need for a three level model over a two level model. Saturated models were formulated with the natural logarithm of time, dummies for FbT, FbTP and NOT, second order interactions between feedback and NOT and third order interactions with time. Both intercept and slope were random. Non-significant predictors (p-out > 0.05) were removed until a parsimonious model was reached, that did not significantly differ from the saturated model. Effect sizes were computed using Equation 1, in which the difference between the estimate at time point t and the baseline OQ-45 score was divided by the baseline OQ-45 standard deviation. Baseline differences were analyzed using a one-way ANOVA with post-hoc Bonferroni correction.

$$d = \frac{estimate_t - estimate_{baseline}}{sd_{baseline}} \quad \text{(Equation1)}$$

Clinically significant and reliable change (Jacobson & Truax, 1991) were computed using the cut-off score for normal functioning of the Dutch OQ-45 that was available at the start of this study based on preliminary analyses (de Beurs, den Hollander-Gijsman, Buwalda, Trijsburg, & Zitman, 2005). The current cut-off score for the Dutch OQ-45 is 55 (De Jong, et al., 2007), but since feedback was provided based on the cut-off score of 58, in the calculations for clinically significant change 58 rather than 55 was used. End status functioning of patients was determined by the last available OQ-45. Last observation carried forward was used if the OQ-45 from the session immediately preceding treatment termination was not available. Differences in clinical significant and reliable change between conditions were tested using a Chi Square test.

Results

Patients

A number of 475 outpatients met the requirements for inclusion in the study and had at least three administrations. Demographic characteristics and diagnoses in each condition are presented in Table 2. There were no significant baseline differences between conditions on the OQ-45 for all participants. However, after excluding patients with less than three OQ-45 administrations, small differences occurred: within the FbT and FbTP groups, the OQ-45 baseline scores of included patients were somewhat higher than those of the excluded patients ($t(203) = -2.48$; $p = 0.014$ and $t(205) = -3.27$; $p = 0.001$ respectively). This resulted in higher baseline scores for FbTP than the control group in the final sample ($F(2, 472) = 4.41$, $p = 0.013$, especially for long-term therapies.

The median therapy length was 35 weeks, and was used to distinguish between short ($n = 231$) and long-term ($n = 243$) therapies. Long-term therapies included more NOT patients ($\chi^2(1) = 13.52$, $p < 0.001$). Baseline differences for the short and long-term therapy group were not significant, except for age. Patients in the long term therapy group were somewhat older ($M = 43.1$, $SD = 12.2$) than patients in the short term therapy group ($M = 40.1$, $SD = 11.7$), $t\ (472) = 2.71$, $p = 0.007$).

Rate of change

The effect of feedback on outcome was examined in two ways: rate of change (speed of progress) and end state functioning (final outcome). The rate of change refers to the steepness of the slope in the change model and indicates how much faster or slower patients change over time due to the factors investigated. Participants did not complete the OQ-45 on every therapy session, but compliance was reasonably good, given that on average more than half of the administered questionnaires were

completed by the patients (see Table 1).

The analyses began by testing if all three levels were required in the multilevel model. The intraclass correlation for the therapist level was computed on an empty model and had a value of 0.02, meaning that only 2% of the total variance in the data was situated at the therapist level. In the three-level model the slope for therapists level was not significant ($\chi^2(2) = 1.47$, $p = 0.48$). Therefore, the therapist level was dropped from subsequent analyses. Table 3 shows the results of the multilevel models on the effect of feedback on symptom reduction on the OQ-45. Table 4 and Figure 2a, 2b and 2c shows the effect sizes of the feedback in the different models after 26, 35, 52 and 78 weeks of treatment.

Then a model for all therapy lengths was analysed. There was an overall significant small positive effect (according to Cohen's criteria (Cohen, 2002)) of feedback to therapists and patients over time, but contrary to expectations no significant effect of feedback to therapists alone was found. Also, no significant interaction was found

Table 2 Patient characteristics

	NFb		FbT		FbTP		Total	
	n	*% or Mean (SD)*	*n*	*% or Mean (SD)*	*n*	*% or Mean (SD)*	*n*	*% or Mean (SD)*
Female	144	65 %	159	64 %	172	74 %	475	68 %
Age	144	42.3 (11.9)	158	41.6 (11.7)	170	41.2 (12.4)	475	41.7 (12.0)
> High school	140	69 %	157	71 %	166	71 %	463	72 %
Diagnoses	121		128		151		400	
Mood disorder		26 %		21 %		31 %		26 %
Adjustment disorder		17 %		20 %		16 %		18 %
Anxiety disorder		15 %		8 %		9 %		10 %
Relational problems (V-codes)		15 %		11 %		15 %		14 %
Other[1]		26 %		40 %		29 %		32 %
Personality disorder		41 %		38 %		34 %		37 %
Co morbidity within axis 1		46 %		44 %		51 %		47 %
Co morbidity axis 1 and 2		43 %		35 %		32 %		36 %
Baseline OQ-45 score								
Included >=3 administrations[2]	144	65.1 (22.4)	159	69.3 (22.5)	172	72.4 (21.9)	475	69.2 (22.4)
Excluded < 3 administrations	48	68.4 (27.4)	46	59.8 (24.2)	35	59.2 (20.7)	129	62.9 (24.8)
Number of sessions	126	33.5 (40.5)	140	36.0 (56.7)	144	27.5 (17.2)	410	32.3 (41.4)
Number of OQ-45 administrations	144	15.7 (16.6)	159	15.8 (15.2)	172	17.4 (18.0)	475	16.4 (16.7)
% completeness per patient	126	55 (28)	140	54 (26)	144	57 (27)	410	57 (27)
Not On Track	144	15 %	159	21 %	172	19 %	475	18 %
Short (< 35 weeks) term	71	7 %	84	17 %	77	10 %	232	12 %
Long (>35 weeks) term	73	22 %	75	25 %	95	26 %	243	25 %

Note. NFb = No Feedback; FbT= Feedback to Therapist; FbTP = Feedback to Therapist and Patient

[1] Other disorders include: disorders usually first diagnosed in infancy, childhood or adolescence, impulse control disorders, eating disorders, dissociative disorders, sexual disorders, substance-related disorders and psychotic disorders (in order of frequency).

Table 3 Fixed and random effects for change trajectories

	All therapy lengths		Short term therapies (<35 weeks)		Long term therapies (≥35 weeks)	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Short term therapies						
Fixed effects						
Intercept	71.79****	1.09	70.89****	1.57	66.96****	2.55
Time	- 4.10****	0.31	- 5.52****	0.42	- 4.05****	0.40
FbT					7.37*	3.37
FbTP					9.21**	3.45
Time * FbTP	- 1.03*	0.51			- 1.46*	0.73
Time * NOT			13.81****	2.04		
Time * NOT * FbT			- 5.93*	2.35		
Time * NOT * FbTP			- 8.31**	3.09		
Random effects						
Intercept	500.97****	36.38	500.3****	53.44	472.0****	47.29
Slope	24.07****	2.23	17.0****	3.40	22.5****	2.50
Covariance	- 43.02****	7.10	- 29.1***	10.55	- 46.3****	8.61
Residual	113.79****	2.02	113.0****	4.54	113.9****	2.26

Note: NFb = No Feedback; FbT = Feedback to Therapist; FbTP = Feedback to Therapist and Patient; NOT = Not On Track; Time is the natural log of weeks + 1. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4 Estimated effect sizes (Cohen's *d*) of difference between groups

	All cases	Not On Track	
	NFb-FbT vs. FbTP	NFb vs. FbT	NFb vs. FbTP
All therapy lengths			
26 weeks	0.15		
35 weeks	0.16		
52 weeks	0.18		
78 weeks	0.20		
Short-term therapies			
26 weeks		0.84	1.18
35 weeks		0.91	1.28
Long-term therapies			
26 weeks	0.22		
35 weeks	0.24		
52 weeks	0.27		
78 weeks	0.29		

Note: NFb = No Feedback; FbT = Feedback to Therapist; FbTP = Feedback to Therapist and Patient

Figure 2a Effect sizes per group for all therapy lengths

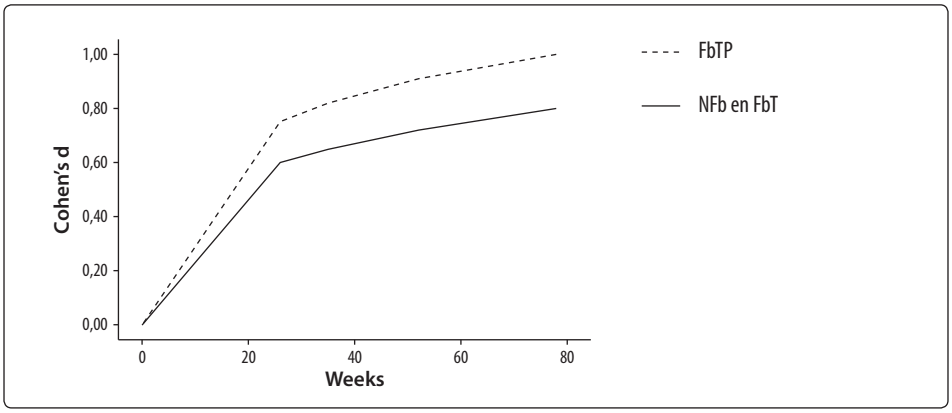


Figure 2b Effect sizes for short-term therapies

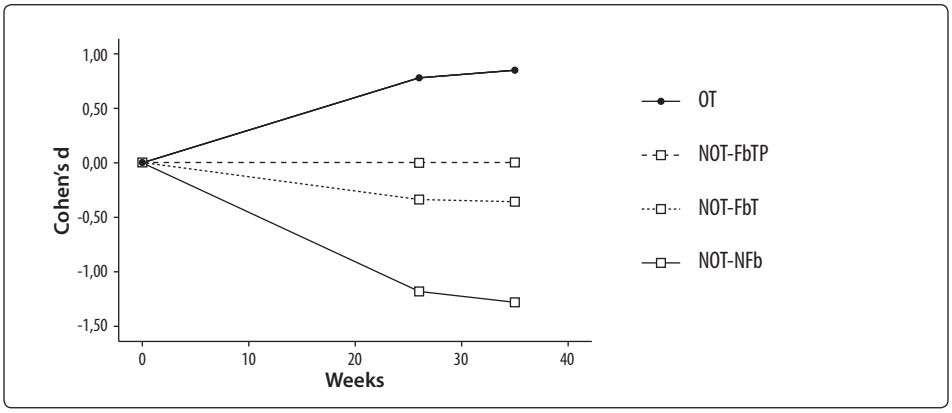
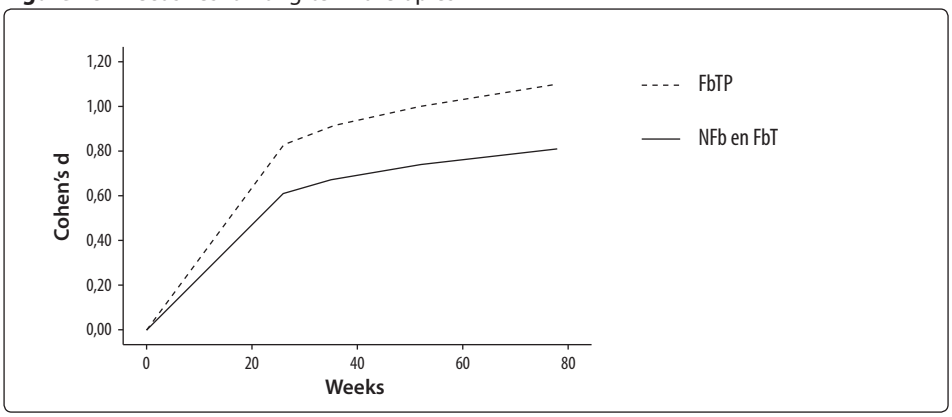


Figure 2c Effect sizes for long-term therapies



between feedback (either FbT or FbTP) and the patient being not on track.

Next, two additional models were analyzed, for short and long-term therapies separately. In short-term therapies there was a significant three-way interaction between time, the status of the patient being not on track, and type of feedback. NOT cases have a negative change over time. In the FbT and FbTP conditions receiving feedback had respectively a large and a very large effect for the NOT cases and was preventive of negative outcomes. The negative effect of being NOT was compensated by receiving feedback, but did not result in positive change. There was no effect of feedback in the OT cases (see Table 3). Feedback to patient and therapists had a small advantage over feedback to therapists alone in the NOT cases, the additional effect sizes are 0.34 at 26 weeks and 0.37 at 35 weeks.

In the long-term therapy group there was a significant difference in OQ-45 scores at baseline for both feedback conditions compared to the NFb control group. Therefore, the baseline OQ-45 scores for the FbT and FbTP groups were included in the model as intercept predictors. The FbTP condition had a favourable small effect on the rate of change, equally of OT and NOT cases (see Table 4).

End state functioning

Table 5 shows the OQ-45 scores at the end of treatment. Although there were no overall significant differences between conditions (χ^2 (6) = 8.01, p = 0.24), there was a trend for the FbTP condition to have the best results: the lowest rate of deteriorated patients (Z = -1.3; p = 0.097) were in this condition. Subgroup analysis of short and long term therapies showed similar results per subgroup, although recovery rates were somewhat better in the long term therapy group than the short term therapy group.

Table 5 Clinical significant and reliable change per condition

	All therapy lengths (n = 475)			Short term therapies (n = 232)			Long term therapies (n = 243)		
	NFb	FbT	FbTP	NFb	FbT	FbTP	NFb	FbT	FbTP
Recovered	37%	38%	43%	32%	30%	35%	41%	48%	50%
Improved	10%	8%	13%	11%	8%	12%	8%	8%	15%
No change	46%	42%	38%	47%	49%	47%	45%	35%	32%
Deteriorated	8%	11%	5%	10%	13%	7%	6%	9%	4%

Note: NFb = No Feedback; FbT = Feedback to Therapist; FbTP = Feedback to Therapist and Patient

Discussion

Summary of results

In this study we aimed to demonstrate the effect of feedback about patient progress to therapists and patients. As anticipated, feedback to both therapists and patients was most effective. The benefits were strongest for cases who were not progressing well in short-term therapies. Feedback provided to the therapist alone was effective of Not On Track (NOT) patients in short-term therapies. No significant effect of being NOT was found in the full sample. In long-term therapies only feedback to therapist and patient was effective. Feedback influenced that rate of change, but did not significantly improve end state functioning.

Short and long term effects

The effects in the short-term group resemble results found by Lambert's group. His group was among the first to study the effect of feedback on patient outcomes and has performed the largest number of studies in the effect of feedback, compared to others. Their studies typically demonstrate that feedback is most effective for NOT cases (Lambert, 2007; Shimokawa, Lambert, & Smart, 2010b). Our overall effect of feedback was 0.15 after 26 weeks and 0.20 after 78 weeks, which is considered a small effect. However, if we look at the effect of feedback for NOT cases in the short-term therapy group – which most resembles Lambert's samples – the effects are very similar to what they found. That is, feedback to therapists and patients had a very large effect in this subgroup of patients and feedback to therapists had a large (after 26 weeks) to very large (after 35 weeks) effect. For these cases the feedback reduced the number of negative outcomes.

For feedback in long-term therapies, Knaup *et al.* (Knaup, et al., 2009) found no significant effect in their meta-analysis. In contrast, we did find a small but significant effect ($d = 0.22$ after 26 weeks and $d = 0.29$ after 78 weeks). Our sample of long-term therapies differed in several ways from the long-term therapies they included in their analysis. They defined long-term effects of feedback as measured between 3 and 12 months after initial assessment, whereas in our long term group treatment duration is much longer. In addition, the majority of the studies they included focused on a more chronic population that includes patients with schizophrenia and chronic (bipolar) depression. Moreover, in three of the five studies the feedback was provided only once or twice. So their long-term group possibly did not include the most effective types of feedback and it may have included a group among whom not much progress might be expected.

These findings raise the question of why feedback seems most effective in NOT cases in short-term therapies but not in long-term therapies. One explanation might be that in longer therapies therapists have more opportunities to identify and correct the negative track. The results show that more positive change trajectories are found in the NOT cases for the long-term therapies. Another issue is that patients who receive long-term therapy are possibly not the same type as patients who receive short-term therapies. Although we did not find relevant differences between patients in long and short-term therapy, patients may differ on unmeasured constructs that are related with the complexity of their complaints. For instance, patients that are not progressing well might have dropped out before 35 weeks in therapy and could be underrepresented in the long-term therapy group. An alternative interpretation is that the OQ-45 might be better at measuring domains that are likely to change in short-term therapies, such as symptoms, than domains that are targeted in long-term therapies, such as character changes.

Therapist versus patient feedback

In the current study, the strongest effect of feedback was found when both therapist and patient received feedback. Our findings may shed light on possible reasons why previous studies on patient feedback have shown mixed results. In the Hawkins *et al.* (Hawkins, et al., 2004) study feedback to patients and therapists outperformed feedback to therapists alone, but the studies by Slade *et al.* (Slade, Lambert, Harmon, Smart, & Bailey, 2008b) and Harmon *et al.* (Harmon, et al., 2005) did not show significant effects of patient and therapist feedback over therapist feedback alone. The overall effect of these three studies resulted in no significant effect (Shimokawa et al., 2010). One of the explanations for the differential effects might be found in different populations. The study by Hawkins et al took place in an outpatient center, whereas the studies of both Harmon *et al.* and Slade *et al.* were done in a university counselling center that provided therapy to students with personal concerns. The outpatient group had more severe patients as well as a more mature group (Shimokawa et al., 2010) and thus, probably resembles our group more than the counselling center sample does.

One could wonder why feedback to therapist and patient shows a more pronounced effect than feedback to therapists alone. Since the therapist is the one providing the therapy, the added effect of providing feedback to patients could be low. There are a couple of explanations that could be viable. For instance, it could be a matter of implementation. The therapist knows that the patient sees the feedback too and this might encourage the therapist to look at the feedback as well. A recent study amongst therapists showed that two major barriers for therapists to look at the feedback were other tasks that demanded attention and not having enough time to look at

the feedback (De Jong, 2012). If the therapist knows that the feedback is not seen by the patient, looking at the feedback might be assigned a lower priority than other administrative tasks. Therapists may also experience resistance to being evaluated (Riemer & Bickman, 2011) and might avoid looking at the feedback as a result. When the patient receives feedback as well, they are not in a position to disregard the feedback, since they know the patient might ask about it.

Another explanation might be that patients are more empowered when they receive feedback about their own progress in therapy. Some of the therapists in our study indicated that providing feedback to patients may have resulted in an increased sense of ownership of their own change process. By receiving the feedback, patients might be more alarmed if there is a lack of progress and might actively discuss this with their therapist and manage their own lack of progress. In that way it may promote communication between patient and therapist.

An alternative explanation is that if patients can track their own progress, they can also manipulate the results and the effect of feedback to therapists and patients might actually be due to a response shift. It is impossible to filter out such an effect, and in our experience some patients will use the feedback to communicate with their therapist through the questionnaires, but the effect of this usually disappears after a few weeks.

Limitations

The current study has some limitations that might influence study results. One of the problems we encountered was that baseline scores were higher for the feedback to therapist and patients group than for the no feedback group. This difference was caused by excluding patients with less than three administrations of the OQ-45 and was most pronounced in the long term therapies. Possibly feedback causes patients with higher complaint levels to stay in the study. We tried to compensate for this problem by adding the baseline scores of the OQ-45 to the multilevel model as a covariate.

A factor that might complicate the generalization of our results is that it is unclear to what extent our sample is selective. Therapists could have made self-selections of patients they approached to participate in the study and we had no way of checking this possibility. Similarly, not all therapists may have reported all patients that refused to participate in the study. The fact that we do not have sufficient insight in the selectiveness of the sample is mainly related to partnering with private practices rather than a single department in a mental health care institution. It was particularly complicated to get information from them on patients for whom the therapists provided treatment in the same time period, but who were not enrolled in the study.

Another issue that needs discussion is our definition of NOT cases. We decided to

use a definition in which a patient needed to have a deteriorated score at least two times. This resulted in relatively low percentages of NOT cases (14-20%), whereas other studies resulted in NOT cases in 20-30% of the cases (Slade, et al., 2008b), and sometimes even up to 50% of the cases (Hawkins, et al., 2004). We chose to have two deteriorations rather than one in order to rule out accidental high scores on the OQ-45 and to ensure that a patient was actually on a negative track.

Finally, our definition of short and long-term therapies has some drawbacks. We divided therapies in two groups post hoc, which resulted in equal group sizes and thus optimal power to detect an effect for both groups, but may have problems when drawing inferences. For instance, it is possible that receiving the feedback had its influence on treatment duration, although we did not find significant differences in treatment duration between the conditions. In addition, it may be that patients not progressing well are over-represented in the long-term therapy group. We did indeed find that there were more NOT in the long-term therapy group, but this could also be due to higher chances for being NOT by having more sessions. Also, since we used treatment duration in weeks, this division does not tell us much about dosage. Short-term treatments might have had a higher density (e.g. weekly sessions rather than bi-weekly).

Implications for practice and research

The current study shows that feedback can be effective in improving the rate of change in outpatient mental health care. Although outcomes were not necessarily better when feedback was provided, progress was achieved faster, which may result in more cost-effective interventions and earlier diminution of suffering. Feedback effects were small in long-term therapy and OT cases. Consistent with previous studies (Lambert et al., 2003), the strongest effects of feedback in our study were found in NOT cases in short-term therapies, so providing feedback is mainly recommended in those cases. It should be noted though that although feedback did seem to prevent a negative treatment course, the effect was not strong enough to result in a positive treatment course for these patients (see Figure 2b). However, previous studies have showed that NOT cases have an increased risk of achieving negative treatment outcomes (Lutz, et al., 2006) and if feedback can help prevent that, it should be considered. The use of an expected recovery curve in the feedback model and using clinical support tools to help the therapist may improve the effect of the feedback, since NOT cases could be identified sooner, based on the deviation of their expected progress (Shimokawa et al., 2010).

Although more studies are emerging on the topic of feedback, there is still much we do not know about the subject yet. There is still little known on how feedback

works in clinical practice and why it improves outcomes in some situations, but does not in others. In addition, most feedback studies have been performed with outpatient adults, and we do not know what the results are in other treatment settings and other groups. Newnham *et al.* (Newnham, Hooke, & Page, 2010) showed for instance that in an acute clinic, feedback was only effective for depressed patients. A recent study by Bickman *et al.* (Bickman, Douglas Kelley, Breda, De Andrade, & Riemer, 2011) in youth mental health care demonstrated differential effects for outcomes measured by clinicians, parents or caregivers and the youth themselves, with the clinicians being most optimistic about these effects and the youth the least. Feedback theory (Riemer & Bickman, 2011) might be able to provide us with a better framework to understand how feedback works. More research is also needed on how therapist and patients use the feedback in therapy. Overall, this study provides us with more knowledge on the effectiveness of feedback to therapists and patients, for short and long-term therapies.

Acknowledgements

This study was supported by a grant from The Netherlands Organisation for Health Research and Development (ZonMW), grant number 94506414. The authors would like to thank Andrew Page for his comments on the first draft of this article.

Chapter 7

General discussion

Main objectives and conclusions

The principal aim of this thesis was to develop an outcome monitoring feedback model for Dutch outpatient mental health care in the Netherlands and to test whether providing feedback to therapists and patients can improve treatment outcomes. First, the psychometric properties of the Dutch translation of the Outcome Questionnaire (OQ-45; Lambert, et al., 2004) were tested and normative samples were collected (Chapter 2). The next step was designing a study in which feedback was provided to therapists. During the design phase of the study the power to detect an effect was attempted to be optimized, by looking for the right proportion of patients within therapists and by anticipating missing data (Chapter 3). Meanwhile, data were collected on patient progress in ten outpatient centers in three different mental health care institutions in both rural and residential areas in the Netherlands. These data were used to predict the functioning of patients at the end of treatment and the speed of recovery (rate of change). The data were analyzed using contemporary statistical techniques, that are flexible in handling missing data (Chapter 4). Finally, two feedback studies were conducted. The first study was a multicenter study in an outpatient mental health care setting and followed patients in their treatment for one year. Feedback was provided to the therapist only and in addition to patient outcomes, therapist characteristics on relevant traits related to feedback effectiveness were studied (see Chapter 5). The second study was conducted in both public outpatient centers and in private practices and contained both short and longer term therapies, up until two years. This study had three treatment conditions, a control group, a group with feedback to therapists alone and a group in which both therapists and patients received feedback about the patient's progress (see Chapter 6).

Cross-cultural validation of the OQ-45

The cross-cultural validation of the OQ-45 results showed that the American and Dutch versions of the OQ-45 are similar when it comes to reliability and validity estimates, but differences in factor structure and normative scores were found. The three-domain structure of the instrument, for which there was no strong evidence in the original version, was slightly better in the Dutch population, but still not satisfactory. Further analyses on the residual correlation matrix, which consisted of the variance that was unexplained by the three factor solution, resulted in two additional factors. The first factor consisted of four items that were in the social role domain. The unexplained variance in this domain was most likely caused by the poor performance of item 14 ('I work/study too much'), a problematic item in the American OQ-45 as well (see Mueller, Lambert, & Burlingame, 1998). The second factor, named Anxiety and Somatic Distress,

was considered a useful addition to the existing scales. Reliability and validity estimates for the ASD factor were promising. This factor might be especially interesting for use by care providers that specialize in anxiety or psychosomatic disorders.

Comparison of normative scores between the American and Dutch populations showed that the Dutch community and clinical samples scored somewhat below their American equivalents. These differences resulted in a cutoff score for the Dutch population (55) of 8 points below the American cutoff point (63). Sensitivity and specificity values were very similar to those of the original version. The reliable change indices were equal (14 points). A marked difference between the American and Dutch normative scores is that in the Dutch population gender differences were found in both the clinical and the community sample. Men had more problems in the social role domain, whereas women showed higher levels of symptom distress as well as anxiety and somatic distress.

The reliability of the subscales and the total scale was adequate in most of the samples. An exception was the internal consistency of the social role domain, which was too low in all three samples, but substantially better when the clinical and community sample were combined. Sensitivity to change was very good and the OQ-45 could effectively discriminate between functional and dysfunctional populations. The concurrent validity showed proper values for the Symptom Distress and Anxiety and Somatic Distress subscales, but less support for the Interpersonal Relations and Social Role subscales. Overall, the Dutch version of the Outcome Questionnaire had sufficient to good psychometric properties.

Power in three-level multilevel models with therapist effects

Multilevel analysis has become increasingly popular for the analysis of longitudinal data in psychotherapy research. Thus far, limited attention had been paid to power analysis in these models. Chapter 3 demonstrates the effects of intraclass correlation, level of randomization, sample size, covariates and drop-out on the power to detect an effect, using data from a routine outcome monitoring study as the basis for simulation studies. A three-level multilevel model was postulated, with therapists at the highest level (level 3), patients within therapists at the middle level (level 2) and measurements within patients at the lowest level (level 1). Results demonstrated that randomization at the patient level was more effective, in terms of power, than randomization at the therapist level. Increasing the number of patients within therapists was shown to be the best way to improve power when randomization took place at the patient level. In the case of randomization at the therapist level, including more therapists was more effective. Increasing the number of measurements per patient did not have a strong effect on power in both randomization designs. In our example, adding gender as a

covariate did not influence the power much. However, our covariate did not have a strong effect and other, more significant covariates may have different effects on power. Drop-out from the study or treatment also did not affect power substantially, although it did reduce power to some extent, especially when drop-out was concentrated at the beginning of the study. Besides power, it is necessary to have appropriate sample sizes at each level to ensure accurate estimation of parameters and standard errors. In some cases this may require larger sample sizes than are necessary for sufficient power. In addition, in order to effectively distinguish between the slope variances at the patient and therapist level, there needs to be a sufficient number of patients per therapist.

Results indicate that in three-level multilevel models larger sample sizes are required than are common in general linear model approaches. This is especially the case in naturalistic data, in which the proportion of variance explained by therapist variance in outcomes is usually larger than in randomized controlled trials (Crits-Christoph, et al., 1991). The larger the portion of variance that is explained by the therapist level (referred to as the intraclass-correlation), the larger the sample size needed. Providing feedback to therapists on their patients' progress may reduce variance in outcomes between therapists, since therapists that have more negative outcomes are provided with the opportunity to adapt their treatments based on the feedback.

Risk factors for negative outcomes

Since one of the main objectives of feedback is to prevent negative outcomes, it would be useful to know which factors are associated with negative outcomes. In Chapter 4 we aimed to predict the risk for negative treatment outcomes at the end of treatment using Classification And Regression Trees (CART) and for rate of change using multilevel modeling. A common problem in finding predictors of outcomes is that naturalistic databases are used and those usually have missing data. Both CART and multilevel analysis are flexible in handling missing data: CART for missing values on the predictor variables and multilevel models for missing values on the dependent variable. Multiple imputation was used to impute missing data in predictor variables in the multilevel analysis.

Fifty-one per cent of the patients in our sample (n = 1540) improved and had scores on the OQ-45 outside the clinical range at the end of treatment. In the CART analyses we found that patients with relatively low pre-treatment scores for symptom distress, and patients with high education have a better chance of favorable outcomes. An extended model, with more nodes (branches) to the regression tree, showed the complexity of the relation between predictors and outcome and showed how pre-treatment expectancies, social role problems and GAF scores and the working alliance at the beginning of treatment (Task subscale) interacted in different ways to predict

negative outcomes at the end of treatment. The multilevel analyses showed that initial severity, the working alliance (Task or Goal subscale) and GAF score were significant predictors for the rate of change in patients. In the complete case sample, having a mood or adjustment disorder as main diagnosis had a positive relationship with the rate of change, whereas in the imputed sample previous treatment, having comorbid Axis I disorders and having a personality disorder as main diagnosis had a negative relationship with the rate of change. The model based on the multiply imputed data was considered the most reliable model, and further analyses were computed only for this model. The CART models and multilevel models differed in their sensitivity to detect negative outcomes. The first CART model had high sensitivity, but low specificity, whereas the multilevel model had high specificity and low sensitivity. The multilevel model was good at picking up deterioration, but not at identifying the no change group. The extended CART model had the best balance between sensitivity and specificity.

Effect of feedback

The effect of feedback on outcome was investigated in two randomized clinical trials. In the first study (Chapter 5), we aimed to research the efficacy of 'simple' (no warning signals or expected recovery curves) feedback compared to no feedback. The largest effects of feedback have been found in models that have expected treatment recovery curves and warning systems for patients that are deviating too much from the expected course. However, most outcome monitoring feedback systems do not have these features and the effectiveness of those systems has been studied insufficiently (Marshall, Haywood, & Fitzpatrick, 2006). Patients (n = 413) were randomly assigned to a no-feedback control group or the feedback condition. Patients that were not progressing well in therapy, so called 'not on track' cases (NOT), were identified post-hoc based on experiencing reliable deterioration in the course of treatment and feedback was expected to be especially effective for them. Contrary to our expectations, for the full sample of therapists no additional beneficial effect of feedback was found and there was no significant interaction between feedback and patients being not on track (NOT). However, in NOT cases a positive significant effect was found when therapists indicated that they used the feedback.

In the second study (Chapter 6) we aimed to demonstrate the additional effect of feedback to therapists and patients. Patients were randomly assigned to three conditions: no feedback, feedback to the therapist alone and feedback to both patient and therapist. Feedback was provided without expected recovery curves, but therapists did get feedback messages that suggested that a patient had deteriorated or not changed. As anticipated, feedback to both therapists and patients was most

effective. Subgroup analyses were performed for short-term (less than 35 weeks of therapy) and long-term (35 weeks of therapy or longer) treatment. The benefits were strongest for NOT cases in short-term therapies. Feedback provided to the therapist alone was also effective for NOT patients in short-term therapies. In long-term therapies only feedback to therapist and patient was effective. In general, feedback influenced the rate of change, but did not significantly improve end state functioning.

Therapist effects

Characteristics of the therapists and the way in which they use feedback may play a central role in the effectiveness of feedback. After all, if therapists do not use feedback constructively, it is unlikely that outcomes will improve. Several characteristics of the therapists that might influence the effectiveness of feedback were studied (Chapter 5). Feedback is more likely to be accepted if it comes from a source that has credibility and has personal relevance to the receiver (Claiborn & Goodyear, 2005). This concept is referred to as perceived validity. Another factor that seems important in acceptance is feedback orientation (Herold & Fedor, 2003; Herold, Parsons, & Rensvold, 1996). External feedback propensity reflects the preference for externally mediated feedback as well as greater faith in such information than in what one can self-generate, whereas internal feedback propensity reflects preference for internally generated feedback as well as the tendency to reconcile differences between internal and external feedback in the direction of internally generated information (Herold & Fedor, 2003). Self-efficacy is another characteristic that influences the feedback process. It refers to a person's beliefs concerning his or her ability to successfully perform a given task or behavior (Bandura, 1977). The commitment to use the feedback in therapy might also be an important factor. Australian research showed that 44% of therapists thought outcome monitoring was a waste of time (Aoun, Pennebaker, & Janca, 2002) and two-third of the therapists was not willing to use the monitoring feedback, not even if it would lead to demonstrably better outcomes (Walter, Cleary, & Rey, 1998).

The results of the study showed that therapists variables moderated the effectiveness of feedback. Therapists with a high internal feedback propensity, who are more likely to trust their own opinion than feedback from external sources, had patients with a slower rate of change than therapists with a low internal feedback propensity, whereas therapists who were more committed to use the feedback at the beginning of the study had patients who progressed faster. Both findings occurred regardless of whether therapists actually received feedback, which suggests that therapists with an open attitude towards getting feedback reach faster progress with their patients. Strangely though, when therapists with a high commitment to use the feedback actually received feedback, this slowed down the rate of change in

their patients. There also was a positive effect of self-efficacy. Patients in the feedback condition who had therapists with higher self-efficacy progressed quicker in therapy than patients of therapist with lower self-efficacy or patients of therapists that did not receive feedback. No effect was found for external feedback propensity and perceived validity. Therapists were more likely to use the feedback if they were more committed to use the feedback at the start of the study and if they were female.

Theoretical and methodological considerations

Symptoms versus functioning

There has been discussion on whether symptom reduction is the best outcome measure to study the effect of psychosocial interventions. Gladis, Gosch, Dishuk and Crits-Christoph (1999) state that a fundamental problem with a symptom-focused approach is that it is based on a narrow, outdated notion of health and disease. Although symptom relief is a major goal of treatment efforts, there are many reasons to expand outcome assessment to include other aspects of clinical progress. The OQ-45, which was used as the outcome measure in the studies in this thesis, measures symptoms as well as social functioning in work and interpersonal relationships and is broader than symptoms alone. However, analyses were on the level of the total score and outcomes were not assessed separately on the different domains of the OQ-45. Using an outcome instrument that assesses multiple domains has its implication on the effect sizes found. Outcome measures that are tailored to specific complaints (for instance the BDI for depression) usually have the highest effect sizes (Lee, Jones, Goodman, & Heyman, 2005). Moreover, since therapy length is under pressure by insurers and is often limited to a certain number of sessions, and studies show that change on symptoms usually occurs before social functioning improves (Howard, Lueger, Maling, & Martinovich, 1993; Stulz & Lutz, 2007) it is likely that a clinically significant improvement in social functioning may not have taken place yet by the end of therapy. A recent (internal) analysis of data collected as part of routine outcome monitoring in our outpatient psychotherapy centers supported this hypothesis and showed that of the patients entering treatment with problems on interpersonal and work functioning, around 70% leaves treatment unchanged on these domains (De Jong & Mooij, unpublished).

Self-report versus other perspectives

Gold and Stricker (2011) state that: "exclusive reliance on self-observation and self-report by the patient may be an ineffective and unreliable method for assessing

psychotherapy. " (p. 1098). Patients may exaggerate or diminish their complaints on a questionnaire in a need to please or rebel against the therapist or by cause of another motivational conflict (Gold & Stricker, 2011). This effect might be more pronounced if patients receive feedback themselves. We saw this phenomenon in a study we recently performed in a inpatient long-term psychotherapy setting for patients with personality disorder. Preliminary data-analysis showed that in the first few weeks, the feedback to patients and therapists group had on average higher dysfunctioning than the feedback to therapists only and no feedback control group. This effect faded after a few weeks (De Jong, Segaar, Busschbach, & Timman, in preparation). An alternative example is that patients sometimes may score higher close to discharge, if they are anxious to end therapy, or lower if they want to end it. Another issue is that patients may not always have insight in their own thoughts and feelings. People often attempt to block out unwanted thoughts and feelings and mental processes in the mind are for a large part implicit and inaccessible (Wilson & Dunn, 2004). Although patient reports have their limitations, the patient perspective in assessing level of symptoms and functioning remains valuable, but should be evaluated in the light of its limitations.
Other perspectives may provide valuable information on the patients' functioning as well. A common perspective in outcome monitoring is that of the therapists, especially in patients that are not capable of completing self-report instruments or in patient populations that are prone to biased reports, such as addiction and forensic populations. Where patients may experience response bias, therapists may be subject to observer bias. For instance, therapists were found to have inaccurate ideas about how symptoms interrelated and this affected their judgments (e.g. Lewis, 1991). The study by Hannan et al. (2005) that was discussed in Chapter 1, demonstrated that therapists tend to overestimate treatment success and underestimate treatment failure. However, making a prediction over time may require different mental processes than assessing functioning of the patients at a certain moment in time. Recent studies on fast and frugal heuristics in decision making demonstrate that with the help of simple decision rules, clinical judgment can outperform statistical predictions (e.g. Katsikopoulos, Pachur, Machery, & Wallin, 2008). Comparisons of measurements from the therapist and patient perspective attest that correlations are usually low and that measurements from the therapist perspective shows more positive effects (Trauer, 2010). It seems that both perspectives have their advantages and disadvantage and provide a different type of information.

Alternative perspectives might be the patients' family and friends, employer or society. Kazdin suggests to include measures in the outcome evaluation that are generally accepted to be of critical importance in everyday life (e.g. days missed from work or amount of arrests) for the population under examination (Kazdin, 2003). Ideally, multiple perspectives are chosen, but collecting information from multiple

sources, especially family members or friends, can be time consuming and costly. Therefore, most outcome measurement systems tend to focus on either the therapist or the patient.

Definition of treatment success and failure

In Chapter 1 we stated that the definition of treatment success and failure is complicated and each definition has its drawbacks. Or, as one reviewer put it: "Defining negative outcomes is a tricky business, that might affect the results.". Although we tried to stay close to existing literature on the topic in our definition of negative outcomes, we did deviate from it to some extent as well. For instance, we did not consider people functioning in the normal range as negative outcomes, even if they had not changed in functioning through the course of therapy. Others have considered the full 'no change' group as negative outcomes (e.g. Lutz, et al., 2006). Whether functioning in the normal range is a good criterion for treatment success remains the question. Some people in the normal population may experience marked symptomatology, but do not seek treatment for it.

An alternative criterion for treatment success is that a patient is mentally healthy when finishing treatment. The World Health Organization (WHO) defines mental health as "A state of well-being in which the individual realises his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community"(WHO, 2001). In outcome research mental health is seldom used as a criterion for treatment success. Traditionally, researchers tended to focus on patients being no longer being mentally *ill*. One reason that mental health is so seldom used as criterion for successful treatment may be that the concept of mental health is complicated to measure and there is no generally accepted theory of what mental health should encompass. Some models of mental health are available (e.g. Jacobson & Greenley, 2001; e.g.Taylor & Brown, 1988), but these have not been adopted on a larger scale.

Cultural equivalence

The cultural equivalence between the Dutch and American versions of the OQ-45 were discussed in Chapter 2. There seemed to be both differences and similarities between the instruments. The most striking difference was found in the normative samples: Dutch respondents tended to report lower scores in dysfunctioning than American respondents. The measurements reported in Chapter 2 were all prior to the start of treatment for the patient group, but the discussion of cultural equivalence can be extended to the topic of change patterns and the use of feedback. Almost all data on expected treatment recovery

curves and feedback has been collected in the United States. Little is known about the cultural equivalence of change patterns and even less about cultural differences in how therapists respond to feedback. Do patients from different countries change in the same way, or are there differences to be expected? Although Dutch respondents score lower than American respondents, outcomes seem similar in the US and the Netherlands. The care system is quite different as well, which makes it hard to disentangle the language and cultural factor on the one hand and the characteristics of the care system and the patients within it on the other hand. The same holds for therapists: the training for therapists is different in the US than in the Netherlands, which may result in different attitudes towards feedback. More research in this area might provide better insight in this issue. In addition, our own sample of patients has its limitations as well. The vast majority of our patient and therapist groups were Caucasian and not much is known on how other cultural subgroups in the Netherlands are responding to therapy.

Implications for clinical practice

Providing outcome monitoring feedback to whom?

Feedback is a complicated process and many factors may influence its effectiveness. Our results show that providing feedback on outcomes in clinical practice can be effective, but is not equally effective in all cases. Results from Chapter 6 suggest that feedback to therapists and patients both might be more effective than feedback to therapists alone. Outcome monitoring feedback may also not be equally effective in all treatment settings: the largest effects were found in short-term therapies (up to 35 weeks). A recent (yet unpublished) study in crisis care patients showed that patients that got feedback were worse off than the no feedback control group (J.J.M. Dekker, personal communication). Getting feedback that you have severe problems and that these are not improving can be demoralizing for patients. The same may be true for patients that are not expected to improve in symptoms, like patients with severe mental disorders or patients that are seeking therapy for personal growth and life issues rather than reducing symptoms. Feedback is probably most effective in improving outcomes if progress is possible but not achieved at that moment. In other patient groups it may be helpful in promoting patient-therapist communication (Carlier, et al., 2010), but may not be helpful in improving outcomes.

Besides feedback to patients and therapists, feedback on outcomes can also be provided at higher levels of the care organization, in which case it contains information on outcomes at the group level. No research has yet been performed on the effect of this type of feedback on outcomes or implementation, but in our experience it can be very stimulating for teams and organizations.

Characteristics of the feedback

Important characteristics of the feedback are the timing and frequency of measuring outcomes and providing feedback and the valence and content of the feedback. Research shows that providing feedback immediately is more effective than delayed feedback (Slade, Lambert, Harmon, Smart, & Bailey, 2008). Deciding on the frequency of measuring outcomes and providing feedback is more complicated. No research has been done on the subject, but common sense tells us that the more frequently outcome is measured, the more likely it is that we are able to detect negative progress early on and the more effective the feedback may be. However, studies (including ours) also show that feedback is most effective for patients that are not progressing well in treatment and that is only a small portion of the patients that are in therapy. Measuring all patients frequently can be time consuming and expensive. Finding a tradeoff between time invested in measuring and preventing negative outcomes must be found. Using the prediction models from Chapter 4 might be instrumental in that. Patients that are at risk for negative functioning at the end of treatment or have negative expected treatment recovery curves could be measured more intensively, whereas patients that are expected to progress well might get a lighter measurement schedule. It should be noted that we have found models with many predictor variables, which may not always be practical to use in clinical practice, since it requires that values on all predictor variables are known. This is often not the case; especially the early treatment predictors (expectancies and early working alliance) are often not available for all patients, which was one of the reasons we encountered so many missing variables in our study.

An alternative option, if one does not have access to prediction models for patients at risk, is to measure all patients frequently (every session) for the first 3-5 sessions of therapy to get an idea of the direction the progress is taking. Research shows that early change is highly predictive of outcome at the end of therapy (e.g. Haas, Hill, Lambert, & Morrell, 2002; Lambert, 2005; Lutz, Stulz, & Kock, 2009), so if patients do not improve in the beginning of therapy, that could be an indication to monitor their progress more closely. In the Netherlands, many outcome monitoring systems in outpatient settings measure outcomes not more than once every three months. Depending on how much therapy is provided in that period, that may not be frequent enough to detect negative outcomes early on.

Feedback to the therapist does not necessarily have to be provided every time outcome is measured in patients. In the study in Chapter 5 we measured outcome every session for the first five sessions of therapy, but provided feedback to the therapists at sessions 1, 3 and 5. Especially when outcome is measured on a session by session basis, providing feedback to therapists every session could be too much, especially in longer

treatments, where providing feedback to therapists may have a reduced effectiveness (see Chapter 6). One model could be to provide feedback only when patients are not progressing well and do not report actively to the therapist if the patient is on track. However, this may demotivate therapists, as they will only get negative feedback in that case. Another option is to provide feedback regularly (e.g. every five sessions) if patients are on track and more often if the patient is not progressing well. Riemer and Bickman (2011) propose a hierarchical system in which one brief feedback signal is provided to therapist (e.g. red or green), and more information is provided only if the patient is not progressing well.

Studies show that feedback is more effective if is more specific (Kluger & DeNisi, 1996). This is not surprising, after all just getting the message that the patient is not progressing well might not be very helpful. It may trigger further inquiry and promote a discussion with the patient on the lack of progress, but it would be better if more specific information were available. Lambert has implemented an instrument that assesses the patients' motivation, social support and alliance with the therapist, as well as events that happened in the patients' life that might help explain why the patient is not doing well. In addition to the assessment, suggestions are given on what interventions might be helpful to improve problems in these areas. A recent meta-analysis showed that using the clinical support tools – as this system is called– has superior outcomes compared to feedback on outcome alone (Shimokawa, Lambert, & Smart, 2010).

In this thesis we have focused on measuring outcomes, but other factors might be useful to monitor as well. Therapists often prefer adding process measures to the outcome monitoring (Riemer & Bickman, 2011). Process measures that have been used include the working alliance, expectancies and motivation. Results on how monitoring these concepts influence outcomes has been scarce and inconsistent. For instance, one study that monitored the working alliance besides outcome showed a positive overall effect on outcome (Reese, Norsworthy, & Rowlands, 2009), but no effect of monitoring the alliance was found in other studies (Crits-Christoph, et al., 2010).

What applies to feedback at the individual level, probably also applies to feedback at the organizational level. Other than individual patient feedback, team level feedback is usually on the outcome of completed cases and has the objective to learn from outcomes on an aggregated level and inform treatment policy. It is usually provided periodically (e.g. once a year). Teams are most likely more willing to accept feedback from a source they perceive as valid and if the information is presented in a clear and easy way. It can be complicated to compare different teams, especially if they differ in patient populations and in the level of implementation of outcome monitoring. It is expected that in organizations feedback on outcome will be most effective for 'not on track' teams - teams that perform below the standard - since they will be motivated

to make changes to improve outcomes. An important question is what should be considered the standard? In the Netherlands, a national benchmark for mental health care aims to provide such a standard, although the discussion on whether we are trying to compare apples and oranges will probably continue, as any comparison has its limitations. By applying statistical case mix correction or providing case mix based comparison groups, the quality of such a comparison can be improved. An interesting alternative approach has been developed by Lambert, who uses prediction models to compute the expected outcomes for a mental health care organization. In that way, the institution can be benchmarked against their own expected outcomes (Hansen, Lambert & Forman, 2002; Hansen & Lambert, 2003).

Therapist and organizational factors

The study in Chapter 5 shows that traits of the therapist may moderate the effect of feedback. Gold and Stricker (2011) indicate that therapists have their 'own needs to avoid the perception of failure' and getting feedback about cases that do not progress well in therapy might cause resistance in the therapists. This may especially be the case for therapists with low self-efficacy (see chapter 5). A recent exploratory study showed that therapists with a high external feedback propensity and who perceived the outcome monitoring feedback as more valid, had a more positive attitude towards outcome monitoring feedback (De Jong, 2012). In Chapter 5 it was demonstrated that having a positive attitude towards the feedback before implementation of outcome monitoring, was predictive of actual use of the feedback.

Getting negative feedback about one's patients can be unpleasant. This unpleasant feeling is referred to in the professional literature as cognitive dissonance and is theorized to be the driving mechanism in feedback. By experiencing cognitive dissonance people get motivated to change their behavior (for instance change their treatment strategy). However, changing behavior is not the only option. As the original experiments by Festinger demonstrated, changing one's cognitions is another possibility to reduce cognitive dissonance (Festinger, 1957). Through the process of causal attribution, therapists may attribute the reasons for lack of progress in the patient outside themselves (e.g. 'it is part of the complaints of this patient'), which may reduce the effectiveness of feedback. On the other hand, therapists that hold themselves accountable and get negative feedback on a regular basis may be at risk for burnout (Riemer & Bickman, 2011). Although traits in the therapists are not within the influence of someone who wants to implement outcome monitoring, it is important to understand the dynamics within the therapists and discuss these with them.

Beside personality traits and external causal attribution, other barriers may be present, that may cause therapists to disregard the feedback. A survey amongst

therapists showed that 'other tasks that asked for attention' (40%), not having enough time (21%) and having trouble interpreting the feedback (21%) were frequent barriers for therapists to use the feedback (De Jong, 2012). Organizational factors that may be barriers to look at the feedback are high administrative pressure or full caseloads. Another factor is accountability. Riemer and Bickman (2011) state that if therapists are not hold accountable for using the feedback, implementation may have little impact. Creating external pressure will increase the likelihood that the therapist will have attention for the feedback.

Future research

Methodological innovations

Although many prediction models that aim to predict outcome in psychosocial interventions are available, it remains complicated to predict negative outcomes. More specifically, the patients that show no (reliable) change seem most complicated to predict. Patients that deteriorate can be predicted reasonably well, but model performance for the no change group is much worse (Finch, Lambert, & Schaalje, 2001; Lutz, et al., 2006). Better prediction models are needed and increasingly, statistical techniques from other disciplines are applied in our field in order to increase prediction precision. One technique that is gaining popularity in psychotherapy research is the use of latent variable models, such as growth mixture modeling. In these models, homogeneous change patterns are identified and based on the characteristics of the change patterns group membership is predicted (e.g. Stulz, Lutz, Leach, Lucock, & Barkham, 2007). This approach uses a reverse method compared to traditional regression models and may provide new insights in which patients are likely to improve and deteriorate.

Many of the techniques used in predicting outcomes have been regression based techniques in which the relation between predictor variables and outcome is assumed to be linear. The CART analysis in Chapter 4 showed that linearity may not always be correct. Kendler (2008) states that we need to start looking at more complex interactions in explanatory models for psychiatric illness and apply models that consider predictors at different levels – inside and macro, within and outside the individual – to better understand what risk factors are relevant in psychiatry. Over the last decade, new applications of statistical techniques have been introduced in clinical psychology that allow for more complex interactions, including state-space models that allow for dynamic individual change patterns (e.g. Fisher, Newman, & Molenaar, 2011) and case-based time-series analysis, that uses bootstrapping as a benchmark for individual change (e.g. Borckardt, et al., 2008).

Most of the methods mentioned above use outcome measures at the scale level, but repeatedly completing questionnaires can place a considerable burden on patients. A promising approach is the application of computerized adaptive testing. Adaptive testing involves the administration of a questionnaire in such a way that an optimal amount of information if obtained in a minimal amount of time. Based on the responses of the patients, only relevant questions that are necessary to determine the severity of the patients complaints are administered, which typically leads to a reduction of items of 50% (Weiss & Kingsbury, 1984). An important initiative in this area is a joint initiative working on building a Patient-Reported Outcomes Measurement Information System (PROMIS; http://www.nihpromis.org). The aim of this network is to develop a large bank of items that measure patient-reported outcomes and allow efficient assessment in clinical research of a wide range of diseases, using state of the art scientific techniques. The outcome measures in PROMIS are constructed according to item-response theory, which has several advantages over classical test-theory (Embretson & Reise, 2000).

Feedback studies

Most of the studies on the effectiveness of feedback have been performed in the United States, using outpatients in university based clinics and counseling centers. Often, patients were students and staff at the university were the study was performed (e.g. Harmon, et al., 2007; Lambert, et al., 2001; Reese, et al., 2009; Slade, et al., 2008; Whipple, et al., 2003). Therapists often were faculty at the university and in that sense used to exposure to research in their clinical work. Our studies were performed in the real world and have demonstrated that implementing feedback may be more complex in those settings. More research should be done in settings outside universities and with other patient groups than outpatients. Some work has been done in inpatient settings (Berking, Orth, & Lutz, 2006; Newnham, Hooke, & Page, 2010) and with substance abuse counseling (Crits-Christoph, et al.).

Another line of research that is necessary is optimizing the feedback. In our studies, we did not use expected treatment recovery curves. So far, Lambert's group is the only one that we know of that has incorporated prediction models in the feedback. More research should be done evaluating the use of prediction models as a benchmark for patients' progress. Especially the combination of the CART and multilevel model (see Chapter 4), which is a new approach, would be interesting to use as a base for feedback. In addition, the clinical support tools should be validated – there is no research supporting the premise that feedback on the alliance, social support and motivation has a specific effect in these domains. An expansion of the clinical support tools could include information based on multidisciplinary treatment guidelines.

Our study (Chapter 5) was among the first studies to research the effect of recipient characteristics in feedback on mental health care. The results demonstrated that therapist characteristics are relevant and more research in this area is needed. Therapist characteristics that might be interesting to study include attribution style, locus of control, personality traits of the therapists and emotional stability. In addition, it would be important to get more insight in the dynamics of how feedback works for whom. Are there therapists that perform worse if they are provided with feedback? Is feedback helpful for all patients?

Another line of research that would be interesting is to combine feedback studies with basic research. For instance, one problem with feedback is that it does not seem to have a learning effect in therapists. This issue could be studied in an experimental design, and is currently being studied by the research group of Andrew Page (personal communication). Another experiment would be to try and manipulate the therapists characteristics by training specific feedback related characteristics. Finding an optimal way to present the feedback to therapists could also be tested in small experimental studies.

Finally, for the further development of feedback, it is crucial that the premises of feedback theory are tested in a clinical context, since most theories originate from social and organizational psychology. Feedback effects are considered context specific and currently the contextualized feedback intervention theory (CFIT; Riemer & Bickman, 2011) is the only available theory that focuses on clinical practice. CFIT is complex and for the largest part untested, therefore alternative theoretical models could be explored as a basis to generate new hypotheses about how feedback works in clinical practice.

Concluding thoughts

Feedback is a complicated process and there are many things we still do not know about it, especially on how it works in clinical practice. This thesis demonstrated that providing outcome monitoring feedback to therapists and patients certainly does not improve outcomes under all circumstances. The metaphor of learning archery blindfolded by Sapyta et al (2005) was used earlier to illustrate why therapists may need outcome monitoring feedback about their patients' progress. They stated that "without direct feedback on how their clients are progressing, clinicians are essentially wearing a blindfold while shooting at a target" (p. 152). This – somewhat bold – statement may be true, but the question remains whether outcome monitoring feedback provides the kind of feedback that is necessary for mastering "hitting the target" and whether it is the only source of feedback. In addition, targets may change during the course of therapy.

People have sometimes asked me if we should consider stopping with outcome monitoring, considering that effects on outcome are smaller than anticipated and the effort to implement it can be large. I do not think that we should. Although effects are small now, with new developments in research and better implementation into clinical practice effects may very well increase. Even with relatively small effects on outcome, and feedback mainly being effective for a portion of patients that are not progressing well, providing feedback on patients' progress can still be useful. As long as patients leave therapy deteriorated or unchanged and there is a chance for change, we owe it to our patients to give it our best shot.

Appendices

A

Appendix A: Equations used to compute power plots

Notation

Level 1	repeated measures	subscript i	n_1 measurements per person
Level 2	patients	subscript j	n_2 persons per therapist
Level 3	therapists	subscript k	n_3 therapists
cond	treatment condition (-1 = control, +1 = experimental)		

Optimality criterion

As optimality criterion we use the variance of the estimator of the interaction effect between treatment and time. If such an interaction exists, then the growth rate differs across treatment conditions. Only for very simple designs an explicit formula for this variance can be derived:

- all subjects measured on the same occasions
- no drop-out; no intermittently missed observations
- number of patients per therapist is fixed
- within each therapist: equal number of patients per treatment condition (randomization at patient level)
- equal number of therapists per treatment condition (randomization at therapist level)

Multilevel model with three levels, randomization at therapist level

We assume that there are $n_3/2$ therapists per treatment condition.

Level-1 model

$$outcome_{ijk} = \beta_{0jk} + \beta_{1jk} time_{ijk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma_e^2)$$

Level-2 model

$$\beta_{0jk} = \gamma_{00k} + u_{0jk} \qquad u_{0jk} \sim N(0, \sigma_{u0}^2)$$
$$\beta_{1jk} = \gamma_{10k} + u_{1jk} \qquad u_{1jk} \sim N(0, \sigma_{u1}^2)$$

Level-3 model

$$\gamma_{00k} = \delta_{000} + \delta_{001} cond_j + v_{0k} \qquad v_{0k} \sim N(0, \sigma_{v0}^2)$$
$$\gamma_{10k} = \delta_{100} + \delta_{101} cond_j + v_{1k} \qquad v_{1k} \sim N(0, \sigma_{v1}^2)$$

Composite model

$$outcome_{ijk} = \delta_{000} + \delta_{100} time_{ijk} + \delta_{001} cond_{jk} + \beta_{101} time_{ijk} cond_{jk} + v_{0k} + v_{1k} time_{ijk} + u_{0jk} + u_{1jk} time_{ijk} + e_{ijk}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma_e^2 + n_1 \sigma_{u1}^2 s_x^2 + n_1 n_2 \sigma_{v1}^2 s_x^2}{n_1 n_2 n_3 s_x^2}$$

Note that $s_x^2 = \frac{1}{n_1} \sum_{j=1}^{n_1} (time - \overline{time})^2$ is the biased variance of the time points.

The standardized effect size is:

$$\delta = \frac{\beta_3}{\sqrt{\sigma_{u1}^2 + \sigma_{v1}^2}}$$

Multilevel model with three levels, randomization at patient level

Assuming that the treatment effect and interaction effect treatment*time *does not vary across therapists*. We assume that for each therapist there are $n_2/2$ patients per treatment condition.

Level-1 model

$$outcome_{ijk} = \beta_{0jk} + \beta_{1jk} time_{ijk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma_e^2)$$

Level-2 model

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01} cond_{jk} + u_{0jk} \qquad u_{0jk} \sim N(0, \sigma_{u0}^2)$$
$$\beta_{1jk} = \gamma_{10k} + \gamma_{11} cond_{jk} + u_{1jk} \qquad u_{1jk} \sim N(0, \sigma_{u1}^2)$$

Level-3 model

$$\gamma_{00k} = \delta_{000} + v_{0k} \qquad v_{0k} \sim N(0, \sigma_{v0}^2)$$
$$\gamma_{10k} = \delta_{100} + v_{1k} \qquad v_{1k} \sim N(0, \sigma_{v1}^2)$$
$$\gamma_{01k} = \delta_{010}$$
$$\gamma_{11k} = \delta_{110}$$

Combined model

$$outcome_{ijk} = \delta_{000} + \delta_{100} time_{ijk} + \delta_{010} cond_{jk} + \delta_{110} time_{ijk} cond_{jk} + v_{0jk} + v_{1jk} time_{ijk} + u_{0jk} + u_{1jk} time_{ijk} + e_{ijk}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma_e^2 + n_1 \sigma_{u1}^2 s_x^2}{n_1 n_2 n_3 s_x^2}$$

Multilevel model with three levels, randomization at patient level

Again, we assume that for each therapist there are $n_2/2$ patients per treatment condition. Now assume that there are differences between therapists in rate of change (time, treatment, treatment*time interaction have random effects). This changes the level-three model:

Level-3 model

$$\gamma_{00k} = \delta_{000} + v_{0k} \qquad v_{0k} \sim N(0, \sigma^2_{v0})$$
$$\gamma_{10k} = \delta_{100} + v_{1k} \qquad v_{1k} \sim N(0, \sigma^2_{v1})$$
$$\gamma_{01k} = \delta_{010} + v_{2k} \qquad v_{2k} \sim N(0, \sigma^2_{v2})$$
$$\gamma_{11k} = \delta_{110} + v_{3k} \qquad v_{3k} \sim N(0, \sigma^2_{v3})$$

Combined model

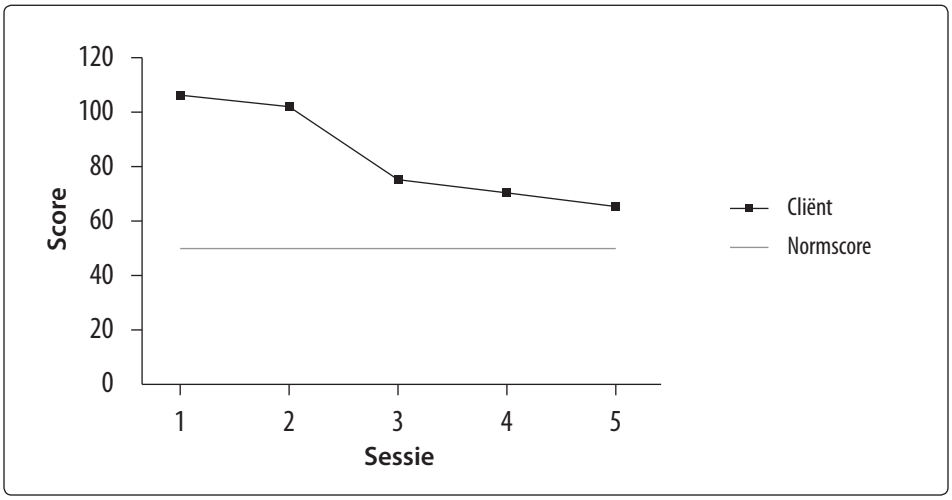
$$outcome_{ijk} = \delta_{000} + \delta_{100} time_{ijk} + \delta_{010} cond_{jk} + \delta_{110} time_{ijk} cond_{jk} + v_{0k}$$
$$+ v_{1k} time_{ijk} + v_{2k} cond_{jk} + v_{3k} time_{ijk} cond_{jk} + u_{0jk} + u_{1jk} time_{ijk} + e_{ijk}$$

$$\mathrm{var}(\hat{\beta}_3) = \frac{\sigma^2_e + n_1 \sigma^2_{u1} s^2_x + n_1 n_2 \sigma^2_{v3} s^2_x}{n_1 n_2 n_3 s^2_x}$$

Appendix B: Example of the feedback used in chapter 5

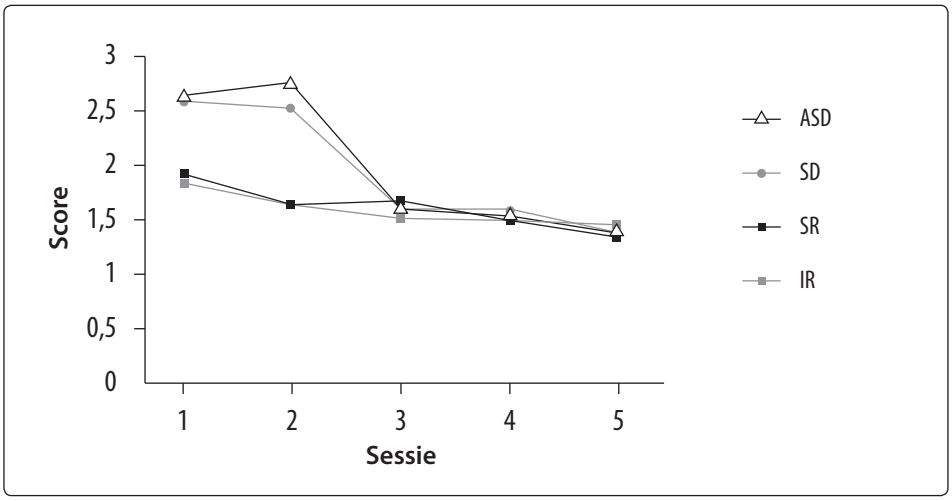
Feedback	
Naam:	Voorbeeld
Geboortedatum:	04-02-1980
Sessienr:	5
Behandelaar:	Behandelaar, X

Totaalscore OQ-45 (grafiek)



Totaalscore OQ-45 (tabel)	
Huidige score:	65
Beginscore:	105
Verandering	40
Status:	Betrouwbaar verbeterd

Gemiddelde schaalscore per item (grafiek)

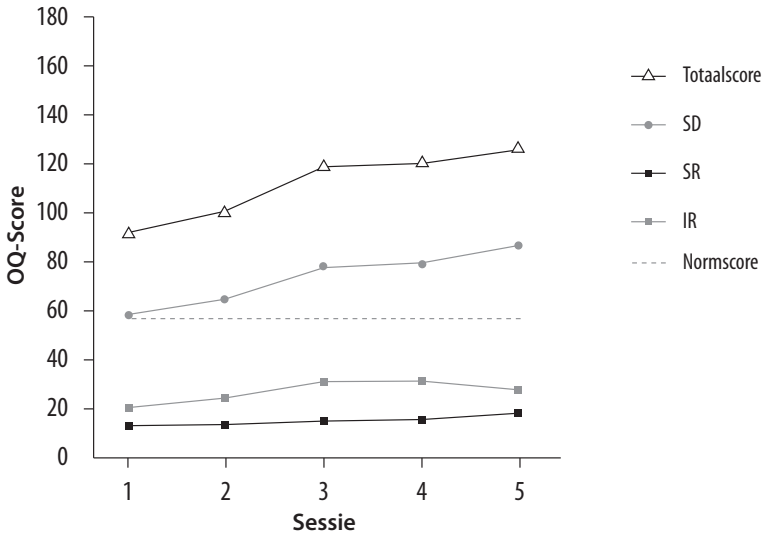


Schaalscores (tabel)			
	Max. score	Begin score	Huidige score
Symptomatische Distress (SD)	100	66	35
Angst en Somatische Distress (ASD)	52	35	18
Interpersoonlijke Relaties (IR)	27	19	16
Sociale Rol (SR)	36	17	12

Risicovragen	
8. Suïcide - Ik denk erover om een einde aan mijn leven te maken	Soms
26. Alcohol/drugs - Ik erger mij aan mensen die kritiek hebben op mijn drinken of drugsgebruik	Zelden
44. Agressie - Ik ben zo kwaad op het werk / op school dat ik iets kan doen waarvan ik spijt zou kunnen krijgen	Zelden

Appendix B: Example of the feedback used in chapter 6

Feedbackbericht naar aanleiding van uw laatste sessie
Laatste sessie: 5 De evaluatie laat zien dat uw cliënt ten tijde van de vorige sessie veel last had van klachten en problemen. Uw cliënt voelt zich slechter dan in het begin van behandeling. Uw cliënt heeft een goede kans om nog verder van de behandeling te profiteren.



Totaalscore	De totaalscore toont u het verloop van u klachten gebaseerd op de 'Outcome Questionnaire'- vragenlijsten die u voorafgaand aan elke therapiesessie invult.
De ernst van de klachten (Symptom distress)	Bij het onderdeel klachten gaat het vooral om klachten van depressie en angst, maar ook over drank- of drugsmisbruik en -afhankelijkheid.
Functioneren in relaties	(Interpersonal relations): Bij het onderdeel relaties gaat het om de omgang met uw partner, andere gezinsleden, familieleden en/of vrienden.
Maatschappelijk functioneren	(Social role): Het onderdeel maatschappelijk functioneren geeft aan hoe het gaat op uw werk, met uw opleiding, of met uw huishoudelijk werk. Problemen op het werk, een hoge werkdruk verslechterd werk of slechte opleidingsresultaten, leiden tot een hoge score op deze schaal.

Uw cliënt heeft bij de laatste invulling van de OQ-45 een verhoogde score aangegeven op één of meer 'kritische items'.
Deze duiden op problemen waarvoor bijzondere aandacht gewenst kan zijn. Als u overweegt het probleem met uw cliënt of bijvoorbeeld met een collega te bespreken, kunnen de volgende punten van belang zijn:
1. Een hogere score duidt op een ernstiger situatie. 0 = nooit 1 = zelden 2 = soms 3 = regelmatig 4 = bijna altijd
2. Uw cliënt kan zich hebben vergist bij het invoeren.
3. Uw cliënt zou kunnen veronderstellen dat de informatie geen deel uitmaakt van de feedback.

Onderwerp		**Item**	**Score**
⚑	Suïcidaliteit	Ik denk erover om een einde aan mijn leven te maken.	3
⚑	Middelen-misbruik	Na zwaar gedronken te hebben, moet ik de volgende morgen weer drinken om op gang te komen.	2
⚑	Middelen-misbruik	Ik erger me aan mensen die kritiek hebben op mijn drinken (of drugsgebruik).	3
⚑	Middelen-misbruik	Ik heb moeilijkheden op het werk/op school door mijn drinken of drugsgebruik.	1
⚑	Agressie op het werk	Ik ben zo kwaad op het werk/op school dat ik iets kan doen, waarvan ik spijt zou kunnen krijgen.	3

Dutch summary

S

Doelstelling

Het hoofddoel van dit proefschrift was om een feedbackmodel voor *routine outcome monitoring* (ROM) te ontwikkelen voor ambulante kortdurende psychiatrische en psychotherapeutische behandelingen in Nederland. Om dit doel te bereiken moest een aantal stappen worden gezet. De eerste stap was om de kwaliteit van de Nederlandse vertaling van de Outcome Questionnaire (OQ-45; Lambert et al., 2004) te onderzoeken (Hoofdstuk 2). Deze vragenlijst wordt gebruikt als uitkomstmaat in de onderzoeken in de rest van het proefschrift. De tweede stap was om te bepalen wat gemiddeld gezien het behandelverloop is van patiënten en welke factoren geassocieerd zijn met een verhoogd risico op negatieve behandeluitkomsten (Hoofdstuk 4). Hiervoor werden gegevens verzameld bij drie GGZ-instellingen. Er werd gebruik gemaakt van geavanceerde statistische technieken om negatieve behandeluitkomsten te voorspellen. De derde stap was het onderzoeken van de effectiviteit van het geven van feedback over het behandelverloop. Hiervoor werden twee onderzoeken gedaan. In elk onderzoek was de feedback anders vormgegeven. Eerst werd bekeken hoe groot de steekproeven in de onderzoeken moesten zijn om een effect te kunnen vinden (*power*) (Hoofdstuk 3). In het eerste feedbackonderzoek werd alleen aan behandelaars feedback gegeven. Dit werd vergeleken met geen feedback (Hoofdstuk 5). In het tweede onderzoek werd feedback aan behandelaars vergeleken met feedback aan zowel behandelaars als patiënten en geen feedback (Hoofdstuk 6). Verder werd onderzocht of eigenschappen van behandelaars invloed hebben op de mate waarin behandelaars gebruik maken van de feedback en op de effectiviteit van de feedback (Hoofdstuk 5).

Cross-culturele validatie van de OQ-45

De cross-culturele validatie van de OQ-45 liet zien dat de Amerikaanse en Nederlandse versies van de OQ-45 vergelijkbaar waren in betrouwbaarheid en validiteit, maar er werden verschillen gevonden in de factorstructuur en de normen. Voor de drie-factor structuur van de OQ-45 met de subschalen Symptomatische Distress (SD), Interpersoonlijke Relaties (IR) en Sociale Rol (SR) was geen sterk bewijs in de Amerikaanse versie. In de Nederlandse versie paste deze structuur beter, maar was nog steeds niet voldoende goed. Verdere analyses leidden tot het toevoegen van een extra factor, die aanvullend op de drie oorspronkelijke factoren gebruikt kan worden. Deze factor, Angst en Somatische Distress (ASD) genoemd, is vooral interessant voor het meten van uitkomsten bij patiënten met angst- of psychosomatische stoornissen. Een vergelijking van de Amerikaanse en Nederlandse normen liet zien dat de Nederlandse normale populatie en poliklinische populatie wat lager scoren

dan Amerikaanse populaties. Deze verschillen leidden ertoe dat de grenswaarde voor normaal functioneren bij de Nederlandse populatie (55) acht punten lager ligt dan de Amerikaanse grenswaarde. De sensitiviteit en specificiteit waren vergelijkbaar met die van de originele Amerikaanse versie. De reliable change index (een maat voor statistisch betrouwbare verandering) was gelijk voor de twee versies (14 punten). Een verschil tussen de Amerikaanse en Nederlandse normen was dat er in de Nederlandse populatie sekseverschillen gevonden werden. Mannen hadden meer problemen op de sociale rol schaal, terwijl vrouwen een hoger klachtenniveau hadden op de schaal symptomatische distress en de schaal angst- en somatische distress.

De betrouwbaarheid van de subschalen en totaalscore was voldoende tot goed, met uitzondering van de sociale rol schaal. De gevoeligheid voor verandering was uitstekend en de OQ-45 kon goed onderscheiden tussen de normale en disfunctionele populaties. De concurrente validiteit van de schalen voor symptomatische distress en angst- en somatische distress was goed, maar was minder sterk voor de schalen interpersoonlijke relaties en sociale rol. Samenvattend kan gezegd worden dat de Nederlandse versie van de OQ-45 voldoende tot goede psychometrische kenmerken heeft.

Power in drie-level multilevel modellen met behandelaareffecten

Multilevel analyse, een vorm van regressieanalyse met geneste data waarbij patiënten bijvoorbeeld genest zijn binnen behandelaars, is de laatste jaren steeds populairder geworden als methode om longitudinale data te analyseren. Dit komt mede doordat multilevel analyse zo flexibel om kan gaan met ontbrekende metingen. Tot op heden is er maar beperkt aandacht besteed aan poweranalyse in deze modellen. Hoofdstuk 3 laat de effecten zien van de intraclass correlatie coëfficiënt (ICC), het niveau van randomiseren, de steekproefgrootte, covariaten en drop-out op de *power* om een effect te vinden. Data uit routine outcome monitoring werden gebruikt als basis voor simulatieonderzoek. Er werd uitgegaan van een drie-level model, met behandelaars op het hoogste niveau (level 3), patiënten binnen behandelaars op het middelste niveau (level 2) en metingen binnen patiënten op het laagste niveau (level 1). De resultaten lieten zien dat het effectiever voor de power was om patiënten willekeurig aan condities toe te wijzen (*randomisatie*) dan behandelaars. Het aantal patiënten vergroten was de beste manier om de power te verhogen wanneer de randomisatie op patiëntniveau plaatsvond. Bij randomisatie op behandelaarniveau was het toevoegen van meer behandelaars effectiever. Het vergroten van het aantal metingen per patiënt had weinig effect op de power in beide onderzoeksdesigns, net als het toevoegen van een covariaat. Drop-out had eveneens geen groot effect op de power, hoewel het de power wel enigszins verlaagde, vooral wanneer de drop-out in het begin van het onderzoek geconcentreerd was.

Risicofactoren voor negatieve behandeluitkomsten

Omdat het voorkomen van negatieve uitkomsten een van de belangrijkste doelen van feedback aan behandelaars is, is het nuttig om te weten welke factoren geassocieerd worden met negatieve behandeluitkomsten. In hoofdstuk 4 was het doel om het risico op negatieve uitkomsten aan het einde van de behandeling te voorspellen. Uitkomsten werden op twee manieren bekeken: de snelheid van veranderen met multilevel analyse en het functioneren van de patiënt aan het einde van de behandeling met Classification And Regression Trees (CART). CART is een vorm van regressieanalyse waarbij automatisch gezocht wordt naar interacties tussen de voorspellers en resulteert in een voorspellingsmodel met een boomstructuur.

Er werden gegevens over het behandelverloop verzameld bij drie GGZ-instellingen. Daarbij werd geen feedback gegeven aan behandelaars of patiënten. In totaal scoorde 51% van de patiënten in de steekproef (n = 1540) aan het einde van de behandeling vergelijkbaar met de normale populatie en kan derhalve als verbeterd worden beschouwd. In de CART analyse vonden we dat patiënten met relatief lage scores op de OQ-45 voor de start van de behandeling en patiënten met een hoger opleidingsniveau betere kansen hadden op positieve behandeluitkomsten. Een uitgebreider CART model liet zien hoe verwachtingen van patiënten over de uitkomst van de behandeling, scores op de sociale rol schaal, scores op de Global Assessment of Functioning (GAF) en de werkalliantie in de beginfase van de behandeling (Taak schaal) op verschillende manieren met elkaar interacteerden. In de multilevel analyse waren de werkalliantie in de beginfase van de behandeling (Taak of Doel subschaal), comorbiditeit op As I, het hebben van een persoonlijkheidsstoornis als hoofddiagnose en de GAF score voorspellend voor de snelheid van verandering bij patiënten.

Er waren verschillen tussen de CART modellen en de multilevel modellen in hun gevoeligheid om negatieve uitkomsten te detecteren. Het eerste CART model (3 takken) had een hoge sensitiviteit, maar een lage specificiteit, terwijl het multilevel model een hoge specificiteit had maar een lage sensitiviteit. Het uitgebreidere CART model (10 takken) had de beste balans tussen sensitiviteit en specificiteit.

Effect van feedback

Het effect van feedback op de uitkomst van de behandeling werd onderzocht in twee gecontroleerde gerandomiseerde onderzoeken. In het eerste onderzoek (Hoofdstuk 5) was het doel om de effectiviteit van 'eenvoudige' feedback (zonder waarschuwingssignalen of verwachte respons curves) te onderzoeken ten opzichte van geen feedback. In de literatuur zijn de grootste effecten van feedback gevonden bij het gebruik van meer complexe feedback modellen waarbij op basis van een statistisch

model een behandelverloop werd voorspeld voor de patiënt, waar het werkelijke behandelverloop tegen afgezet werd en waarbij de behandelaar een signaal kreeg als de patiënt teveel afweek van het verwachte beloop. De meeste outcome monitoring feedback systemen hebben deze functionaliteit echter niet en de effectiviteit van deze meer 'eenvoudige' systemen is onvoldoende onderzocht, terwijl deze wel op grote schaal gebruikt worden. Patiënten (n = 413) werden willekeurig toegewezen aan een controlegroep zonder feedback of de feedback conditie. Er werd verwacht dat de feedback vooral effectief zou zijn voor patiënten die niet goed vooruit gingen in de behandeling, de zogenaamde 'not on track' (NOT) patiënten. Tegen de verwachtingen in was de feedback echter niet effectiever dan geen feedback en er was geen interactie tussen feedback en NOT zijn voor patiënten. Echter, wanneer behandelaars aangaven dat ze de feedback actief gebruikt hadden in de behandeling, had het geven van feedback aan de behandelaar een significant positief effect voor NOT patiënten.

In het tweede onderzoek (Hoofdstuk 6) was het doel om het effect te laten zien van feedback aan zowel patiënt als behandelaar. Patiënten werden willekeurig toegewezen aan drie condities: geen feedback, feedback aan alleen de behandelaar en feedback aan behandelaar en patiënt. Feedback werd gegeven zonder voorspellingsmodel met verwacht behandelverloop, maar behandelaars kregen wel feedback berichten die aangaven of de patiënt verslechterd was of onvoldoende verandering had doorgemaakt. Er werden aparte analyses gedaan voor korte behandelingen (korter dan 35 weken) en langere behandelingen (35 weken of langer). Het effect van feedback was het sterkste bij NOT patiënten in korte behandelingen, wanneer feedback werd gegeven aan zowel de behandelaar als de patiënt. Feedback aan alleen de behandelaar was ook effectief bij NOT patiënten in korte behandelingen, maar in langere behandelingen was alleen feedback aan behandelaar en patiënt effectief. De feedback had vooral effect op de snelheid waarmee patiënten vooruitgingen in de behandeling, maar had geen significant effect op het niveau van functioneren van patiënten aan het einde van de behandeling.

Behandelaar effecten

Kenmerken van de behandelaars en de manier waarop zij feedback gebruiken spelen mogelijk een belangrijke rol in de effectiviteit van feedback. Immers, als behandelaars de feedback niet constructief gebruiken, is het onwaarschijnlijk dat de feedback de uitkomsten verbetert. Er werden diverse kenmerken van de behandelaar die potentieel van invloed zijn op de effectiviteit van de feedback onderzocht (Hoofdstuk 5). Feedback wordt door behandelaars eerder geaccepteerd wanneer het van een betrouwbare bron komt, die persoonlijke relevantie heeft voor de ontvanger van de feedback. Dit concept wordt *perceived validity* genoemd. Een andere factor die van

belang is bij de acceptatie van feedback is de feedback voorkeur. Mensen met een interne feedback voorkeur vertrouwen meer op hun eigen oordeel dan op feedback uit een externe bron, terwijl mensen met een externe feedback voorkeur zelf actief feedback zoeken, ook bijvoorbeeld bij collega's. Vertrouwen in de eigen competenties (*self-efficacy*) is een andere factor die van invloed is op het feedbackproces. Het verwijst naar het vertrouwen dat iemand heeft om een bepaalde taak succesvol uit te voeren. De motivatie om de feedback te gebruiken in de behandeling is ook een belangrijke factor, omdat bekend is dat behandelaars niet altijd gemotiveerd zijn om feedback te gebruiken en deze dan waarschijnlijk niet effectief is.

De resultaten van ons onderzoek lieten zien dat kenmerken van de behandelaar de effectiviteit van feedback modereren. Behandelaars met een hoge interne feedback voorkeur hadden patiënten die minder snel vooruit gingen in de behandeling dan patiënten van behandelaars met een lage interne feedback voorkeur. Behandelaars die voor aanvang van het onderzoek gemotiveerder waren om de feedback in de behandeling te gaan gebruiken, hadden patiënten die sneller vooruit gingen. Beide effecten traden op ongeacht of de behandelaar feedback ontving of niet en lijken te suggereren dat behandelaars met een open houding ten aanzien van feedback snellere vooruitgang boeken bij hun patiënten. Opvallend genoeg ging de verandering langzamer bij patiënten waarvan de behandelaar een hoge motivatie had om de feedback te gebruiken wanneer deze de feedback ook daadwerkelijk ontving. Verder gingen patiënten in de feedback conditie waarvan de behandelaar meer vertrouwen had in de eigen competenties sneller vooruit in de behandeling dan patiënten van behandelaars die minder vertrouwen hadden of dan patiënten in de controle conditie. Er werd geen effect gevonden van een externe feedback voorkeur en *perceived validity*. Behandelaars waren meer geneigd om de feedback daadwerkelijk te gebruiken in de behandeling als ze voorafgaand aan het onderzoek hoger scoorden op de motivatievragenlijst en vrouw waren.

Consequenties voor de praktijk

Routine outcome monitoring en het geven van feedback aan behandelaars is momenteel sterk in opkomst in Nederland, mede onder druk van de zorgverzekeraars. Dit proefschrift laat zien dat feedback een ingewikkeld proces is en dat vele factoren van invloed zijn op de effectiviteit ervan. De resultaten laten zien dat feedback geven over het behandelverloop effectief kan zijn, maar niet onder alle omstandigheden. De resultaten in hoofdstuk 6 suggereren dat feedback aan zowel patiënten als behandelaars het meest effectief is.

Feedback is bij voorkeur direct beschikbaar en vaker meten biedt meer mogelijkheden om negatieve uitkomsten te voorkomen. Het kan echter kostbaar

zijn om bij alle patiënten frequent te meten. De voorspellingsmodellen in Hoofdstuk 4 kunnen een leidraad zijn om te bepalen welke patiënten een verhoogd risico hebben op negatieve uitkomsten en deze kunnen dan intensiever gevolgd worden in de behandeling. Bij andere patiënten kan dan minder vaak gemeten worden. Een alternatief is om bij alle patiënten gedurende de eerste sessies elke sessie te meten en daarna te bepalen hoe vaak verder meten noodzakelijk is.

Onderzoek laat zien dat feedback effectiever is wanneer het specifieker is. In het kader daarvan is het nuttig om de uitkomsten uit te breiden met *clinical support tools* (zie Shimokawa, Lambert & Smart, 2010), die aangeven welke problemen er zijn en welke interventies mogelijk behulpzaam zijn, wanneer patiënten *not on track* zijn. *Clinical support tools* bevatten procesmaten, zoals de werkrelatie met de behandelaar en de motivatie van de patiënt en veel behandelaars geven de voorkeur aan dit type informatie boven uitkomstmetingen. Een ander aandachtspunt is dat niet alle behandelaars de feedback gebruiken en dat deze derhalve dan ook niet effectief kan zijn. Het krijgen van negatieve feedback kan onaangenaam zijn en erin resulteren dat de behandelaar de oorzaak voor de negatieve feedback, of het gebrek aan vooruitgang van de patiënt, buiten zichzelf legt, wat de effectiviteit van de feedback zou kunnen verminderen. Er zijn ook andere factoren die ervoor kunnen zorgen dat een behandelaar de feedback niet gebruikt, zoals geen tijd hebben om de feedback te bekijken of moeite hebben om de feedback te interpreteren.

Wat op individueel niveau geldt, is ook te vertalen naar een geaggregeerd niveau, bijvoorbeeld een organisatie of team. Anders dan bij feedback aan behandelaars gaat feedback op teamniveau vaak over afgesloten behandelingen en heeft het meer tot doel om te leren van de totale uitkomsten in een bepaalde periode of om beleid op te baseren. Het is te verwachten dat ook bij teams feedback het meest effectief is bij '*not on track*' teams, oftewel teams die minder goede resultaten halen dan de standaard, maar het is ingewikkeld te bepalen wat de standaard zou moeten zijn. Benchmarken zou daar een oplossing voor kunnen bieden.

References

R

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology, 57*, 785–794.

Addis, M. E., Hatgis, C., Krasnow, A. D., Jacob, K., & Bourne, L. (2004). Effectiveness of cognitive-behavioral treatment for panic disorder versus treatment as usual in a managed care setting. *Journal of Consulting and Clinical Psychology, 72*, 625-635.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*, 341-382.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*, 545-557.

Ambler, G., Omar, R. Z., & Royston, P. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research, 16*, 277-298.

Ambler, G., Omar, R. Z., Royston, P., Kinsman, R., Keogh, B. E., & Taylor, K. M. (2005). Generic, simple risk stratification model for heart valve surgery. *Circulation, 112*, 224-231.

American Group Psychotherapy Association Science to Service Task Force (2007). Practice guidelines for group psychotherapy. Retrieved November 2011, from http://www.agpa.org/guidelines/index.html

Anderson, T., Ogles, B. M., Patterson, C. L., Lambert, M. J., & Vermeersch, D. A. (2009). Therapist effects: Facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology, 65*, 755-768.

Aoun, S., Pennebaker, D., & Janca, A. (2002). Outcome measurement in rural mental health care: A field trial of rooming-in models. *Australian Journal of Rural Health, 10*, 302-307.

APA Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

APA Interdivisional Task Force on Evidence-based Therapy Relationships (2011). Conclusions and recommendations of the interdivisional (APA Divisions 12 & 29) task force on evidence-based therapy relationships, Retrieved November 2011, from http://www.divisionofpsychotherapy.org/continuing-education/task-force-on-evidence-based-therapy-relationships/conclusions-of-the-task-force/

Arrindell, W.A., & Ettema, J.H.M. (1975). *Klachtenlijst (SCL-90)*. Lisse: Swets & Zeitlinger.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191-215.

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology, 47*, 397-415.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist, 65*, 13-20.

Berking, M., Orth, U., & Lutz, W. (2006). How effective is systematic feedback of treatment progress to the therapist? An empirical study in a cognitive-behavioural-oriented inpatient setting. *Zeitschrift für Klinische Psychologie und Psychotherapie: Forschung und Praxis, 35,* 21-29.

Beutler, L. E., Malik, M., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 89-108). New York: Wiley.

Bickman, L. (2008). A measurement feedback system is necessary to improve mental health outcomes. *Journal of American Academy of Child and Adolescent Psychiatry, 47*, 1114-1119.

Bickman, L., Douglas Kelley, S., Breda, C., De Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services, 62,* 1423–1429.

Blais, M. A., Sinclair, S. J., Baity, M. R., Worth, J., Weiss, A. P., Ball, L. A., et al. (2011). Measuring outcomes in adult outpatient psychiatry. *Clinical Psychology and Psychotherapy*. doi: 10.1002/cpp.749

Boelen, P.A., de Keijser, J., & van den Bout, J. (2001). Psychometrische eigenschappen van de Rouw Vragenlijst (RVL). *Gedrag en Gezondheid*, *29*, 172-185.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63*, 77-95.

Borkovec, T. D., & Nau, S. D. (1972). Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry, 3*, 257-260.

Borm, G. F., Melis, R. J. F., Teerenstra, S., & Peer, P. G. M. (2005). Pseudo cluster randomization: a treatment allocation method to minimize contamination and selection bias. *Statistics in Medicine, 24*, 3535-3547.

Bosker, R. J., Snijders, T. A., & Guldemond, H. (1996). User's manual PinT. Program and manual available at http://www.gamma.rug.nl/

Bouman, T.K. (1995). De Agoraphobic Cognitions Questionnaire (ACQ). *Gedragstherapie*, *27*, 69-72.

Bouman, T.K. (1998). De Body Sensation Questionnaire (BSQ). *Gedragstherapie*, *31*, 162-168.

Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Briand, B., Ducharme, G. R., Parache, V., & Mercat-Rommens, C. (2009). A similarity measure to assess the stability of classification trees. *Computational Statistics and Data Analysis, 53*, 1208-1217.

Brom, D., & Kleber, R.J. (1985). De schokverwerkingslijst. *Nederlands Tijdschrift voor Psychologie*, *40*, 164-168.

Burgess, P., Pirkis, J., & Coombs, T. (2006). Do adults in contact with Australia's public sector mental health services get better? *Australia and New Zealand Health Policy, 3*, 9.

Butcher, J., Derksen, J., Sloore, H., & Sirigatti, S. (2003). Objective personality assessment of

people in diverse cultures: European adaptions of the MMPI-2. *Behaviour Research and Therapy*, *41*, 819-840.

Campbell, M. K., Mollison, J., & Grimshaw, J. M. (2001). Cluster trials in implementation research: Estimation of intracluster correlation coefficients and sample size. *Statistics in Medicine, 20*, 391-399.

Carlier, I. V. E., Meuldijk, D., Van Vliet, I. M., Van Fenema, E., Van der Wee, N. J. A., & Zitman, F. G. (2010). Routine outcome monitoring and feedback on physical or mental health status: evidence and theory. *Journal of Evaluation in Clinical Practice*, *18*, 104-110.

Castonguay, L. G., Boswell, J. F., Constantino, M. J., Goldfried, M. R., & Hill, C. E. (2010). Training implications of harmful effects of psychological treatments. *American Psychologist, 65*, 34-49.

Chambless, D.L., Caputo, G.C., Bright, P., & Gallager, R. (1984). Assessment of fear in agoraphobics: The Body Sensations Questionnaire and the Agoraphobic Cognitions Questionnaire. *Journal of Consulting and Clinical Psychology*, *52*, 1090-1097.

Chapman, J.E. (2003). *Reliability and validity of the progress questionnaire: An adaptation of the Outcome Questionnaire*. Philadelphia, PA: Drexel University.

Claiborn, C. D., & Goodyear, R. K. (2005). Feedback in psychotherapy. *Journal of Consulting and Clinical Psychology, 61*, 209-217.

Clarkin, J. F., & Levy, K. N. (2004). The influence of client variables on psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed., pp. 194-226). New York: John Wiley & Sons.

Cohen, J. (2002). A power primer. *Psychological Bulletin, 112*, 155-159.

Constantino, M. J., Arnkoff, D. B., Glass, C. R., Ametrano, R. M., & Smith, J. Z. (2011). Expectations. *Journal of Clinical Psychology, 67*, 184-192.

Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., Carroll, K., Perry, K. (1991). Meta analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1*, 81-91.

Crits-Christoph, P., & Gallop, R. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research, 16*, 178-181.

Crits-Christoph, P., Ring-Kurtz, S., Hamilton, J. L., Lambert, M. J., Gallop, R., McClure, B., et al. (2011). A preliminary study of the effects of individual patient-level feedback in outpatient substance abuse treatment programs. *Journal of Substance Abuse Treatment*, doi: 10.1016/j.jsat.2011.09.003.

Crits-Christoph, P., Ring-Kurtz, S., McClure, B., Temes, C., Kulaga, A., Gallop, R., et al. (2010). A randomized controlled study of a web-based performance improvement system for substance abuse treatment providers. *Journal of Substance Abuse Treatment, 38*, 251-262.

Cronbach, L.J. (1990). *Essentials of psychological testing.* New York, NY: Harper Collins Publishers.

De Beurs, E., van Dyck, R., Marquenie, L.A., Lange, A., & Blonk, R.W.B. (2001). De DASS: Een vragenlijst voor het meten van depressie, angst en stress. *Gedragstherapie*, *34*, 35-53.

De Beurs, E., den Hollander-Gijsman, M., Buwalda, V., Trijsburg, W., & Zitman, F. (2005). De

Outcome Questionnaire (OQ-45): Een meetinstrument voor meer dan alleen psychische klachten. *De Psycholoog, 40*, 53–63.

De Jong, A., & van der Lubbe, P.M. (2001). *Groningse vragenlijst voor Sociaal Gedrag: Zelfbeoordelingsvragenlijsten voor het vaststellen van problemen in het interpersoonlijk functioneren (handleiding)*. Groningen: Rob Giel Onderzoekscentrum.

De Jong, K. (2012). De rol van de behandelaar: de 'vergeten' factor in ROM. *Tijdschrift voor Psychiatrie, 54*, 197-201.

De Jong, K., & Nugter, M.A. (2004). De Outcome Questionnaire: Psychometrische kenmerken van de Nederlandse vertaling. *Nederlands Tijdschrift voor Psychologie*, *59*, 76-79.

De Jong, K., Nugter, M. A., Lambert, M. J., & Burlingame, G. M. (2009). *Handleiding voor afname en scoring van de Outcome Questionnaire (OQ-45)*. Salt Lake City, UT: OQ-45 Measures LLC.

De Jong, K., Nugter, M., Polak, M. G., Wagenborg, J. E., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology and Psychotherapy, 14*, 288-301.

De Jong, K., Segaar, J., Busschbach, J. v., & Timman, R. (in preparation). The effect of feedback in outpatient and inpatient personality disorders: a randomized controled trial.

Derogatis, L.R. (1975). *The brief symptom inventory*. Baltimore, MD: John Hopkins University School of Medicine.

Derogatis, L.R. (1977). *The SCL-90 Manual: Scoring, administration and procedures for the SCL-90.* Baltimore, MD: John Hopkins University School of Medicine.

Devilly, G. J., & Borkovec, T. D. (2000). Psychometric properties of the credibility/expectancy questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry, 31*, 73-86.

Dimidjian, S., & Hollon, S. D. (2010). How would we know if psychotherapy were harmful? *American Psychologist, 65*, 21-33.

Dimidjian, S., & Hollon, S. D. (2011). Introduction. *Cognitive and Behavioral Practice, 18*, 303-305.

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H. (2008). Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology, 64*, 344-354.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London, UK: Arnold Publishers.

Edmunds, M., Frank, R., Hogan, M., McCarty, D., Robinson-Beale, R., & Weisner, C. (Eds.). (1997). *Managing Managed Care: Quality Improvement in Behavioral Health*. Washington, DC: National Academy Press.

Elkin, I., Falconnier, L., Martinovich, Z., & Mahoney, C. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16*, 144-160.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S., Collins, J., et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry, 46*, 971-982.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence

Erlbaum Associates.

Essock, S. M., Drake, R. E., Frank, R. G., & McGuire, T. G. (2003). Randomized controlled trials in evidence based mental health care: Getting the right answer to the right question. *Schizofrenia Bulletin, 29*, 115-123.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J. (2002). Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM. *The British Journal of Psychiatry, 180*, 51-60.

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical outcomes in routine evaluation. *Journal of Mental Health*, *9*, 247-255.

Festinger, L. (1957). *A theory of cognitive dissonance*. Standford, CA: Stanford University Press.

Finch, A. J. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology and Psychotherapy, 8*, 231-242.

Fisher, A. J., Newman, M. G., & Molenaar, P. C. M. (2011). A quantitative method for the analysis of nomothetic relationships between idiographic structures: Dynamic patterns create attractor states for sustained posttreatment change. *Journal of Consulting and Clinical Psychology, 79*, 552-563.

Flaherty, J.A., Gaviria, F.M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A., & Birz, S. (1988). Developing Instruments for cross-cultural psychiatric research. *Journal of Nervous Mental Disorders*, *176*, 257-263.

Garb, H. N. (2005). Clinical Judgment and Decision Making. *Annual Review of Clinical Psychology, 1*, 67-89.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analyisis. *Journal of General Internal Medicine, 22*, 1595-1602.

Gladis, M., Gosch, E. A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 320-331.

Gold, J., & Stricker, G. (2011). Failures in psychodynamic psychotherapy. *Journal of Clinical Psychology, 67*, 1096-1105.

Goldstein, H. (2003). *Multilevel statistical models (3rd ed.)*. London: Edward Arnold.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. (pp. 201-218). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213.

Haas, E., Hill, R. D., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy

maintain treatment gains. *Journal of Clinical Psychology, 58*, 1157-1172.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*, 155-163.

Hansen, N. B., & Lambert, M. J. (2003). An evaluation of the dose-response relationship in naturalistic treatment settings using survival analysis. *Mental Health Services Research, 5*, 1-12.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9*, 329-343.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2003). The psychotherapy dose effect in naturalistic settings revisited: Response to Gray. [Comment/Reply]. *Clinical Psychology: Science and Practice, 10*, 507-508.

Haringsma, R., Engels, G. I., Van der Leeden, R., & Spinhoven, P. (2006). Predictors of response to the coping with depression course for older adults. A field study. *Ageing and Mental Health, 10*, 424-434.

Harmon, C., Hawkins, E. J., Lambert, M. J., Slade, K., & Whipple, J. L. (2005). Improving outcomes for poorly responding clients: The use of clinical support tools and feedback to clients. *Journal of Clinical Psychology, 61*, 175-185.

Harmon, S., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., et al. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research, 17*, 379-392.

Hatfield, D., McCullough, L., Frantz, S. H., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology and Psychotherapy, 17*, 25-32.

Hatfield, D.R., & Ogles, B.M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice*, *35*, 485491.

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K. L., & Tuttle, K. C. (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research, 14*, 308-327.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87.

Hentschel, U. (2005). Therapeutic alliance: The best synthesizer of social influences on the therapeutic situation? On links to other constructs, determinants of its effectiveness, and its role for research in psychotherapy in general. *Psychotherapy Research, 15*, 9-23.

Herold, D. M., & Fedor, D. B. (2003). Individual differences in feedback propensities and training performance. *Human Resource Management Review, 13*, 675-689.

Herold, D. M., Parsons, C. K., & Fedor, D. B. (1997). *Individual feedback propensities and their effects*

on motivation, training success and performance. Atlanta, GA: United States Army Research Institute for the Behavioral and Social Sciences.

Herold, D. M., Parsons, C. K., & Rensvold, R. B. (1996). Individual differences in the generation and processing of performance feedback. *Educational and Psychological Measurement, 56*, 5-25.

Hofstede, G. (2006). Cultural dimensions. Retrieved March, 2006, from http://www.geert-hofstede.com/ hofstede_dimensions.php

Hollenbeck, J. R., & Klein, H. J. (1987). Goal commitment and the goal-setting process: problems, prospects, and proposals for future research. *Journal of Applied Psychology, 72*, 212-220.

Hopko, D. R., Magidson, J. F., & Lejuez, C. W. (2011). Treatment failure in behavior therapy: focus on behavioral activation for depression. *Journal of Clinical Psychology, 67*, 1106-1116.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*, 223-233.

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology, 38*, 139-149.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting and Clinical Psychology, 61*, 678-685.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*, 1059-1064.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Huang, Y., Kotov, R., de Girolamo, G., Preti, A., Angermeyer, M., Benjet, C., et al. (2009). DSM-IV personality disorders in the WHO World Mental Health Surveys. *The British Journal of Psychiatry, 195*, 46-53.

Hüpscher, L. (2007). *Herstelverwachtingen van cliënten met een ambulante behandeling*. Master thesis. Amsterdam, the Netherlands: Vrije Universiteit.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336-352.

Jacobson, N., & Greenley, D. (2001). What is recovery? A conceptual model and explication. *Psychiatric Services, 52*, 482-485.

Jacobson, N.S., Roberts, L.J., Berns, S.B., & McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300-307.

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.

Jelicic, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45*, 1195-1199.

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl to fast and frugal heuristics (and back). *Theory and Psychology, 18*, 443-464.

Kazdin, A. E. (2003). Clinical significance: Measuring whether interventions make a difference. In A. E. Kazdin (Ed.), *Methodological Issues and Strategies in Clinical Research (3rd ed.).* (pp. 691-710): Washington, DC, US: American Psychological Association.

Kazdin, A. E. (2008). Evidence-based treatment and practice: new opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146-159.

Kendler, K. S. (2008). Explanatory models for psychiatric illness. *The American Journal of Psychiatry, 165*, 695-702.

Kim, D. M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16*, 161-172.

Klein, D. N., Schwartz, J. E., Santiago, N. J., Vivian, D., Vocisano, C., Castonguay, L. G., et al. (2003). Therapeutic alliance in depression treatment: Controlling for prior change and patient characteristics. *Journal of Consulting and Clinical Psychology, 71*, 997-1006.

Kline, R.B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: Meta-analysis. *The British Journal of Psychiatry, 195*, 15-22.

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61*, 285-314.

Lambert, M. J. (2005). Early response in psychotherapy: Further evidence for the importance of common factors rather than "placebo effects". *Journal of Clinical Psychology, 61*, 855-869.

Lambert, M. J. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*, 1-14.

Lambert, M. J. (2011). What have we learned about treatment failure in empirically supported treatments? Some suggestions for practice. *Cognitive and Behavioral Practice, 18*, 413-420.

Lambert, M.J., Burlingame, G.M., Umphress, V., Hansen, N.B., Vermeersch, D.A., Clouse, G.C., Christopherson, C., & Burlingame, G.M. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy*, *3*, 249-258.

Lambert, M.J., Hannöver, W., Nisslmüller, K., Richard, M., & Kordy, H. (2002). Fragebogen zum Ergebnis von Psychotherapie: Zur Reliabilitat und Validität der deutschen Übersetzung des Outcome Questionnaire 45.2 (OQ-45.2). *Zeitschrift für Klinische Psychologie und*

Psychotherapie: Forschung und Praxis, *31*, 4046.

Lambert, M.J., Harmon, C., Slade, K., Whipple, J.L., & Hawkins, E.J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*, 165-174.

Lambert, M. J., Huefner, J. C., & Nace, D. K. (1997). The promise and problems of psychotherapy research in a managed care setting. *Psychotherapy Research, 7*, 321-332.

Lambert, M. J., Morton, J. J., Hatfield, D. R., Harmon, C., Hamilton, S., Shimokawa, K. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)* (3 ed.). Wilmington, DE: American Professional Credentialling Services LLC. Zoek versie 2003 in artikelen

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed., pp. 139-193). New York, NY: John Wiley & Sons.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*, 149-164.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice, 10*, 288-301.

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11*, 49-68.

Lampropoulos, G. K. (2011). Failure in psychotherapy: An introduction. *Journal of Clinical Psychology, 67*, 1093-1095.

Laurenceau, J.-P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. *Clinical Psychology Review, 27*, 682-695.

Lee, W., Jones, L., Goodman, R., & Heyman, I. (2005). Broad outcome measures may underestimate effectiveness: An instrument comparison study. *Child and Adolescent Mental Health, 10*, 143-144.

Lewis, G. (1991). Observer bias in the assessment of anxiety and depression. *Social Psychiatry and Psychiatric Epidemiology, 26*, 265-272.

Lewis, R. J. (2000). *An introduction to Classification and Regression Tree (CART) Analysis*. Paper presented at the Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, CA.

Liebowitz, M.R. (1987). Social phobia. *Modern Problems in Pharmacopsychiatry*, *22*, 141-173.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lovibond, S.H., & Lovibond, P.F. (1995). Manual for the Depression Anxiety Stress Scales. Sydney, Australia: Psychology Foundation of Australia.

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. *Journal of Consulting and Clinical Psychology, 69*, 150-158.

Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirically based clinical practice. *Psychotherapy Research, 12*, 251-272.

Lutz, W. (2003). Efficacy, effectiveness, and expected treatment response in psychotherapy. *Journal of Clinical Psychology, 59*, 745-750.

Lutz, W., Lambert, M. J., Harmon, S., Tschitsaz, A., Schurch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration – What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy, 13*, 223-232.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., et al. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology, 73*, 904-913.

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54*, 32-39.

Lutz, W., Lowry, J., Kopta, S., Einstein, D. A., & Howard, K. I. (2001). Prediction of dose-response relations based on patient characteristics. *Journal of Clinical Psychology, 57*, 889-900.

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology, 67*, 571-577.

Lutz, W., Rafaeli, E., Howard, K. I., & Martinovich, Z. (2002). Adaptive modeling of progress in outpatient psychotherapy. *Psychotherapy Research, 12*, 427-443.

Lutz, W., Stulz, N., & Kock, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of Affective Disorders, 118*, 60-68.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92.

Margison, F. R., McGrath, G., Barkham, M., Mellor Clark, J., Audin, K., Connell, J. (2000). Measurement and psychotherapy. *The British Journal of Psychiatry, 177*, 123-130.

Marshall, S., Haywood, K., & Fitzpatrick, R. (2006). Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of Evaluation in Clinical Practice, 12*, 559-568.

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 68*, 438-450.

Meyer, T.J., Miller, M.L., Metzger, R.L., & Borkovec, T.D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*, 487-495.

Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*, 199-208.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129-149.

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable. *The American Statistician, 59*, 72-78.

Moerbeek, M. (2006). Power and money in cluster randomized trials: when is it worth measuring a covariate? *Statistics in Medicine*, *25*, 2607-2617.

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. *Journal of Educational and Behavioral Statistics, 33*, 41-61.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: a confirmatory factor analysis. *Journal of Personality Assessment, 70*, 248-262.

Newnham, E. A., Hooke, G. R., & Page, A. C. (2010). Progress monitoring and feedback in psychiatric care reduces depressive symptoms. *Journal of Affective Disorders, 127*, 139-146.

Nich, C., & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology, 65*, 252-261.

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy, 10*, 361-373.

Orlinsky, D. E., Ronnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: continuity and change. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.). New York, NY: Wiley.

Prigerson, H.G., Kasl, S.V., & Jacobs, S.G. (1997). *The Inventory of Complicated Grief Revised.* Unpublished manuscript.

Raudenbusch, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*, 85-116.

Raudenbusch, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Ravitz, P., McBride, C., & Maunder, R. (2011). Failures in interpersonal psychotherapy (IPT): factors related to treatment resistances. *Journal of Clinical Psychology, 67*, 1129-1139.

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, Research, Practice, Training, 46*, 418-431.

Reise, S. P., & Duan, N. (2003). Design issues in multilevel studies. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: methodological advances, issues and applications* (pp. 285-298). Mahwah, NJ: Lawrence Erlbaum Associates.

Riemer, M., & Bickman, L. (2011). Using program theory to link social psychology and program

evaluation. In M. M. Mark, S. I. Donaldson & B. Campbell (Eds.), *Social Psychology and Evaluation*. New York: Guilford Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B.A., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., & Keller, M.B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QUIDS-C) and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*, 573-583.

Sanavio, E. (1988). Obsessions and compulsions: The Padua Inventory. *Behaviour Research and Therapy*, *26*, 167-177.

Sapyta, J., Riemer, M., & Bickman, L. (2005). Feedback to clinicians: Theory, research, and practice. *Journal of Clinical Psychology, 61*, 145-153.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*, 1-10.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*, 298-311.

Simon, W., Harris, M., & Lambert, M. J. (2011). *The effects of progress and clinical support tools feedback compared to TAU within a hospital-based outpatient clinic*. Paper presented at the Society of Psychotherapy Research, 42nd international meeting.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurence*. New York: Oxford University Press.

Slade, K., Lambert, M. J., Harmon, S., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology and Psychotherapy, 15*, 287-303.

Slee, N., Garnefski, N., Van der Leeden, R., Arensman, E., & Spinhoven, P. (2008). Cognitive–behavioural intervention for self-harm: randomised controlled trial. *British Journal of Psychiatry, 192*, 202-211.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-259.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE Publications.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16*, 173-177.

Spybrook, J., Raudenbush, S. W., Liu, X.-F., Congdon, R., & Martínez, A. (2008). Optimal Design (Version 1.76). Ann Arbor, MI: University of Michigan.

Stirman, S.W., DeRubeis, R. J., Crits-Christoph, P., & Brody, P. E. (2003). Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new

methodology and initial findings. *Journal of Consulting and Clinical Psychology, 71*, 963-972.

Stulz, N., & Lutz, W. (2007). Multidimensional patterns of change in outpatient psychotherapy: The phase model revisited. *Journal of Clinical Psychology, 63*, 817-833.

Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of Consulting and Clinical Psychology, 75*, 864-874.

Stuurgroep ROM ggz (2010). *Visie op ROM in de GGZ*. Amersfoort, the Netherlands: GGZ Nederland.

Tasca, G. A., Balfour, L., Ritchie, K., & Bissada, H. (2007). The relationship between attachment scales and group therapy alliance: Growth differs by treatment type for women with binge-eating disorder. *Group Dynamics: Theory, Research and Practice, 11*, 1-14.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193-210.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research articles: A simplified methodology. Retrieved February, 2006, from http://work-learning.com/effect_sizes.htm

Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines. Retrieved April 2011, from http://www.mayo.edu/hsr/techrpt/61.pdf

Trauer, T. (Ed.). (2010). *Outcome measurement in mental health. Theory and practice.* Cambridge: University Press.

Valenstein, M., Mitchinson, A., Ronis, D. L., Alexander, J. A., Duffy, S. A., Craig, T. J. (2004). Quality indicators and monitoring of mental health services: What do frontline providers think? *American Journal of Psychiatry, 161*, 146-153.

Van Balkom, A.J.L.M., de Beurs, E., Hovens, J.E.J.M., & van Vliet, I.M. (2004). Meetinstrumenten bij angststoornissen. *Tijdschrift voor Psychiatrie*, *10*, 687-692.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*, 681-694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049 - 1064.

Van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity, 32*, 15-29.

Van der Ploeg, E., Mooren, T.T.M., Kleber, R.J., van der Velden, P.G., & Brom, D. (2004). Construct validation of the Dutch versions of the Impact of Events Scale. *Psychological Assessment*, *16*, 16-26.

Van Ginkel, J. R., Sijtsma, K., van der Ark, L., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 17-30.

Van Oppen, P., Hoekstra, R.J., & Emmelkamp, P.M.G. (1995). The structure of obsessive compulsive

symptoms. *Behaviour Research and Therapy*, *33*, 15-23.

Van Rijsoort, S., Vervaeke, G., & Emmelkamp, P. (1999). The Penn State Worry Questionnaire and the Worry Domains Questionnaire: Structure, reliability and validity. *Clinical Psychology and Psychotherapy*, *6*, 297-307.

Van Wijngaarden, B., Schene, A.H., Koeter, M., VazquezBarquero, J.-L., Knudsen, H.-C., Lasalvia, A., McCrone, P., & The EPSILON Study Group. (2000). Caregiving in schizophrenia: Development, internal consistency and reliability of the Involvement Evaluation Questionnaire-European Version: EPSILON Study 4. *British Journal of Psychiatry*, *177* (Suppl. 39), S21-S27.

Vervaeke, G., & Vertommen, H. (1993). De werkalliantie: visies op een bruikbaar concept en de meting ervan. *Tijdschrift voor Psychotherapie, 19*, 2-16.

Vervaeke, G., & Vertommen, H. (1996). De werkalliantievragenlijst (WAV). *Gedragstherapie, 29*, 139-144.

Walter, G., Cleary, M., & Rey, J. M. (1998). Attitudes of mental health personnel towards rating outcome. *Journal of Quality in Clinical Practice, 18*, 109-115.

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16*, 184-187.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5*, 425-433.

Watson, J. C. (2011). Treatment failure in humanistic and experiential psychotherapy. *Journal of Clinical Psychology, 67*, 1117-1128.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weiss, D., & Marmar, C. (1997). The impact of event scale—revised. In J. Wilson & T. Keane (Eds). *Assessing psychological trauma and PTSD* (pp. 399-411). New York, NY: Guilford.

Weissman, M.M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry*, *33*, 1111-1115.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent pscyhotherapy. *Journal of Consulting and Clinical Psychology, 65*, 688-701.

Westen, D. (2005). Patients and treatments in randomized trials are not adequately representative of clinical practice. In J. C. Norcross (Ed.), *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions* (pp. 161-170). Washington, DC: American Psychological Association.

Westra, H. A., Constantino, M. J., & Aviram, A. (2011). The impact of alliance ruptures on client outcome expectations in cognitive behavioral therapy. *Psychotherapy Research, 21*, 472-481.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59-68.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement.

Annual Review of Psychology, 55, 493-518.

Wing, J. K., Beevor, A. S., Curtis, R. H., Park, S. B., Hadden, S., & Burns, A. (1998). Health of the Nation Outcome Scales (HoNOS). Research and development. *The British Journal of Psychiatry, 172*, 11-18.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.* (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Dankwoord

D

Ten eerste wil ik alle patiënten die aan dit onderzoek hebben meegewerkt bedanken. Aan de onderzoeken in dit proefschrift hebben zeer veel instellingen en vrijgevestigde therapeuten bijgedragen en zonder hen had ik dit onderzoek niet kunnen doen. In het bijzonder dank ik de behandelaars en managers bij GGZ Noord-Holland-Noord, GGZ Dijk en Duin, PsyQ Haaglanden en de therapeuten in het Montoring onderzoek. Vele studenten van de Universiteit van Amsterdam, de Vrije Universiteit en Universiteit Leiden hebben met hun master these onderzoek een bijdrage geleverd aan de dataverzameling. Zij waren mijn aanspreekpunt op de onderzoekslocaties en samen waren we een sterk team. Verder ben ik GGZ Noord-Holland-Noord zeer erkentelijk dat zij mij in de gelegenheid hebben gesteld om dit onderzoek uit te voeren binnen en buiten de organisatie. Dank ook aan het Erasmus MC en Jan van Busschbach dat ik de resultaten uit het Monitoring onderzoek mocht opnemen in dit proefschrift.

Een proefschrift schrijven doe je gelukkig niet helemaal alleen. Daarom wil ik alle co-auteurs van de artikelen die in dit proefschrift zijn opgenomen hartelijk danken voor hun bijdrage en de prettige samenwerking. Daarbij wil ik vooral mijn waardering uitspreken voor mijn promotoren Philip Spinhoven en Willem Heiser en mijn co-promotor Annet Nugter. Ik heb veel gehad aan jullie adviezen en ben dankbaar dat jullie me de ruimte hebben gegeven om mijn eigen koers te kiezen. Degenen die mij hebben geholpen bij de complexe statistische technieken in dit proefschrift wil ik ook graag nadrukkelijk noemen. Marike Polak, Rien van der Leeden, Mirjam Moerbeek, Cor Ninaber en Joost van Ginkel, ik was heel blij met jullie geduld met de vele vragen die ik stelde en jullie positieve houding en betrokkenheid bij de artikelen waar we samen aan gewerkt hebben. *Special thanks are due to Michael Lambert and Wolfgang Lutz for their great mentorship. I would also like to express my gratitude towards my fellow-researchers at the Society for Psychotherapy Research, who never fail to inspire me with their interesting work.*

Tot slot wil ik mijn familie, vrienden en collega's bedanken voor de steun die zij mij gegeven hebben tijdens dit omvangrijke project en het begrip dat zij hadden voor mijn afzondering in het afgelopen jaar. In het bijzonder bedank ik daarbij mijn ouders, mijn zus Frea en broer Joeri en natuurlijk mijn man Jochem, zonder wiens onvoorwaardelijke steun ik dit project nooit had willen doen.

Curriculum vitae

CV

Kim de Jong was born on June 8, 1977 in Alkmaar. In 1995 she graduated from the Bertrand Russell College in Krommenie and started studying Psychology at the University of Amsterdam. In 2000 she received her master degree ('met genoegen'), in Clinical Psychology and in Psychological Methods. For her master thesis, she translated the Outcome Questionnaire (OQ-45) into Dutch and collected preliminary data on the psychometric properties of the OQ-45. She also did a combined clinical and research internship at polikliniek Ypenstein in Heiloo (GGZ Noord-Holland-Noord).

In October 2000 she started working as a researcher at GGZ Noord-Holland-Noord. During the first few years she combined this job with teaching at the Hotelschool The Hague (research methods and statistics) and the University of Amsterdam (research practicum). She continued to collect data on the OQ-45 which eventually resulted in writing the manual for the Dutch OQ-45. At GGZ Noord-Holland-Noord she mainly performed health services research, which was used for internal reports to inform management on different aspects of the quality of care. She performed studies in a wide variety of patients and settings, including inpatient crisis care, crisis home care, elderly inpatients, adult and forensic outpatients and severe mental illnesses on a great variety of topics, including patient requests, patient needs, satisfaction, dialectical behavioral therapy and brief solution-focused psychotherapy.

In 2006 she started working on her dissertation part-time, as an external candidate, linked to the department of Clinical, Health and Neuropsychology of Leiden University. The results of her dissertation research are presented in this thesis. Starting September 2008, she visited Michael Lambert's research group at Brigham Young University for 5 months. In October 2009, she took on a part-time job as a researcher at the department of Medical Psychology and Psychotherapy at the Erasmus University Medical Center on the Monitoring Psychotherapy study. The results of that study are included in this thesis. She currently works at both GGZ Noord-Holland-Noord and the Erasmus University Medical Center.

In addition to her regular activities, she has taken part in many additional professional activities, including being the coordinator of the OQ-45 international researchers collaborative, being a student representative in the educational committee of the Dutch-Flemish Postgraduate School in Experimental Psychopathology and she was a member of a national task force on routine outcome monitoring (werkgroep Vergelijkbaarheid, project ROMggz). She is currently an associate editor for the Dutch Journal of Psychotherapy (Tijdschrift voor Psychotherapie). She is also very actively involved in the Society for Psychotherapy Research.