2010-06-30

A Patient-Focused Psychotherapy Quality Assurance System: Meta-Analytic and Multilevel Analytic Review

Kenichi Shimokawa
Brigham Young University - Provo

A Patient-Focused Psychotherapy Quality Assurance System:

Meta-Analytic and Multilevel Analytic Review

Kenichi Shimokawa

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Michael J. Lambert
Bruce N. Carpenter
Scott A. Baldwin
David W. Smart
John Okiishi

Department of Psychology

Brigham Young University

August 2010

ABSTRACT

A Patient-Focused Psychotherapy Quality Assurance System:

Meta-Analytic and Multilevel Analytic Review

Kenichi Shimokawa

Department of Psychology

Doctor of Philosophy

Outcome research has documented worsening among a minority of the patient population (5 to 10%). In this study a psychotherapy quality assurance system intended to enhance outcomes in patients at risk of treatment failure was reviewed through the use of meta-analytic, mega-analytic, and multilevel analytic techniques. A pooled dataset from six major studies conducted at a large university counseling center and a hospital outpatient setting (N = 6151, mean age = 23.3 years, female = 63.2%, Caucasian = 85%) were re-analyzed to examine the effects of progress feedback on patient outcome. In this quality assurance system, the Outcome Questionnaire-45 was routinely administered to patients to monitor their therapeutic progress and was utilized as part of an early alert system to identify patients at risk of treatment failure. Patient progress feedback based on this alert system was provided to clinicians to help them intervene before treatment failure occurred. Intent-to-treat and efficacy analyses of the effects of feedback interventions were conducted to obtain the estimates of effects expected from implementation of this quality assurance system as a policy as well as in clinical trials. Three forms of feedback interventions—integral elements of this quality assurance system—were effective in enhancing treatment outcome, especially for signal alarm patients. Two of the three feedback interventions were also effective in preventing treatment failure (Clinical Support Tools and the provision of patient progress feedback to therapists). The Clinical Support Tool intervention was effective not only in terms of the amount of outcome enhancing effect, but also in the rate of patient recovery. The current state of evidence appears to support the efficacy and effectiveness of feedback interventions in enhancing treatment outcome.

Keywords: treatment outcomes, treatment failure, patient deterioration, feedback, psychotherapy quality assurance

ACKNOWLEDGEMENTS

I am grateful for many individuals who supported me throughout this project. I express my profound gratitude for the great mentor and support Dr. Mike Lambert has been for me, not just for this project, but in many aspects of my career development. His continued confidence in me helped me persevere through this project's undertaking, while his dedication to psychotherapy research inspired me to aspire for better integration of research and practice in psychotherapy. I wish to thank Joe Olsen and Scott A. Baldwin for their advice in the statistical analyses. I would like to express my appreciation to Dr. Olsen for his generous sharing of his time and expertise in managing and analyzing large datasets. I am grateful to Rachel Meibos, Seyed Nejad, Aaron Allred, Yasutaka Morita, and Brennan Atherton for their assistance.

Completion of this project and pursuit of my education would have been simply impossible without my family's support. Words cannot express the gratitude I feel towards my dear wife, Hitomi, for so lovingly and patiently supporting me, and the joy and meaning she gives to my life. My dear children, Hiroto, Miku, and Yuuya have been a great source of joy and encouragement. I am deeply grateful for my parents' support for my education.

This research was supported in part by an endowed chair research grant to Dr. Mike Lambert that funded my research assistantship and other student research assistants for which I am also deeply grateful.

Table of Contents

List of Tables and Figure

(All tables and the figure are presented in Appendix A.)

A Patient-Focused Psychotherapy Quality Assurance System:

Meta-Analytic and Multilevel Analytic Review

In this era of accountability, healthcare systems, including mental healthcare systems, have been placed under a tremendous pressure to demonstrate the effectiveness of their service in bringing about improved patient outcome (Lambert, Bergin, & Garfield, 2004; Reed & Eisman, 2006). Within this context, psychology as a discipline has increased its emphasis on what is commonly referred to as *evidence-based practice in psychology* (EBPP; APA, 2006). However, the notion of EBPP has been conceptualized and practiced in various ways. For example, the past decade saw enthusiasm for identifying particular treatments that work more or less effectively in specific contexts (e.g., disorders) and populations (Chambless et al., 1996, 1998). This emphasis on identifying what has been termed as *empirically supported treatments* (ESTs), is often referred to as the EST movement; however, this movement has raised concerns from many psychologists about the exclusive efforts to develop brief, manualized treatments (APA, 2006). Various groups of psychologists, including those representing several divisions of APA, National Institute of Mental Health, and the Department of Health and Human Service's Substance Abuse and Mental Health Services Administration, embarked on developing frameworks to conceptualize and examine scientifically based practice at both the state and federal levels (e.g., National Institutes of Health, 2002). Within this context, concerns had steadily increased regarding the use and misuse of evidence from psychotherapy research findings, such as insurance companies placing restrictions on both the amount of care and the choice of treatments.

These concerns led to the appointment of the APA Presidential Task Force on Evidence-Based Practice (Task Force) in 2005 to make explicit psychology's stance to consider the "full

range of evidence that policymakers must consider" and its "fundamental commitment to sophisticated evidence-based psychological practice" (APA, 2006, p. 273). In this Task Force's recent report (APA, 2006), the following definition for EBPP was set forth: "*Evidence-based practice in psychology* (EBPP) is the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences" (p. 273; emphasis included in the original text). Although varying views about the conceptualization and practice of EBPP have been raised, including criticisms about the lack of inclusiveness of the APA Presidential Task Force's philosophical stance (e.g., Wendt & Slife, 2006), the Task Force's attempt to recognize a broader range of evidence is to be congratulated.

Regarding the phrase "clinical expertise" in this definition, the Task Force expounded the following (APA, 2006; p. 276-277):

> Clinical expertise also entails the monitoring of patient progress (and of changes in the patient's circumstances—e.g., job loss, major illness) that may suggest the need to adjust the treatment (Lambert, Bergin, & Garfield, 2004). If progress is not proceeding adequately, the psychologist alters or addresses problematic aspects of the treatment (e.g., problems in the therapeutic relationship or in the implementation of the goals of the treatment) as appropriate.

Such practice of monitoring and modifying treatment response has been an important aspect of ensuring quality care for clients, and thus has been a central element of patient-focused research (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert et al., 2003).

The paradigm of patient-focused research advocates systematic evaluation of a patient's response to treatment throughout the course of therapy (Howard, et al. 1996). In this paradigm, the patient is considered the most critical informant of distress and thus the outcome. The

advocates of this approach recommend providing clinicians with the feedback of their client's progress. Such feedback allows the therapist to make treatment decisions based on client distress rather than on fixed treatment protocols and mandated policies regarding treatment length. Based on the patient-focused research paradigm, several quality assurance systems have been developed (Beutler, 2001; Kordy, Hannöver, & Richard, 2001; Lambert, Hansen, & Finch, 2001; Lueger et al., 2001).

An important area of psychotherapy research literature from an EBPP perspective and a serious concern for patient-focused research paradigm is the consistent findings that not all clients derive benefit from psychotherapy, and that some actually get worse in treatment, experiencing an increase of distress (Hansen, Lambert, & Forman, 2002). These patients are often referred to as treatment failures or non-responders (Mash & Hunsley, 1993). Although a causal relationship cannot be easily drawn between psychotherapy and the failing outcome, it should be alarming to a profession that seeks to be of help to its consumers that their "customers" leave their service worse off than when they first came for help (Finch, Lambert, & Schaalje, 2001). While these cases of treatment worsening may comprise only about 10% of total cases, failing cases also impose serious economic implications to persons who did not get the benefit they sought and to the third party payers who endorsed the ineffective treatments. In addition to the economic implications of treatment failures, the sheer number of patients whose quality of life worsens despite receiving psychotherapy places prevention of treatment failure as an important goal of outcome research and practice.

To help prevent treatment failures, Lambert and colleagues (Lambert, Hansen, et al., 2001) developed a quality assurance system to enhance the treatment outcome of the individual patient. This system is established on three major principles: (1) developing a reasonable

estimate of the expected progress of the average patient based on the patient scores on the Outcome Questionnaire-45.2 (OQ; description of this instrument provided below); (2) a data driven process of comparing the actual progress of a patient of interest to expected progress, and (3) the provision of treatment progress information (feedback) to the therapist and case managers. Six major studies have been conducted to evaluate the effects of providing feedback about patient progress and the expected progress trajectory of the patient. These studies have been published elsewhere (Harmon, Lambert, Slade, Smart, & Lutz, 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Slade, Lambert, Harmon, Smart, & Bailey, 2008; Whipple et al., 2003).

The basic rationale behind the concept of providing feedback to clinicians is straightforward. Therapist can be more responsive to patient needs if they know that the patient is not succeeding as intended. As has been repeatedly demonstrated in clinical research, any prediction relying on statistical or actuarial methods tend to fare better than clinical judgment alone (see Ægisdóttir et al., 2006 for a recent major meta-analytic review on this subject). Such actuarial/clinical prediction research can be traced back to the work of Paul Meehl, who reached the same conclusion in 1954 (see Grove, 2005, for an in-depth discussion). This notion is especially salient when clinicians are making predictions about treatment failures. For instance, more recently, Hannan and colleagues (Hannan et al., 2005) have demonstrated how clinicians fare poorly in predicting deteriorated cases in psychotherapy. Hatfield, McCullough, Frantz, and Krieger (2009) found only 32% of therapists recorded patient worsening in their case notes, despite dramatic escalation in their symptoms in the week prior to meeting with their therapist.

The psychological community has increasingly recognized the importance of providing feedback to clinicians regarding their patients' progress. For instance, APA Task Force (2006)

noted that one of the "most pressing research needs" in EBPP includes this particular type of research, which they recapitulated as "providing clinicians with real-time patient feedback to benchmark progress in treatment and clinical support tools to adjust treatment as needed (p.278)."

In the following sections, a brief description of the feedback system developed by Lambert's team, including the overview of the designs of the past major studies, a brief summary of the past developments in Lambert and his colleague's feedback studies, and the rationale, purposes, methodology, and limitations of the present study are described.

Feedback System

Defining outcome. In each of the six studies examined in this study, patients' psychological dysfunction was measured by the Outcome Questionnaire-45.2 (OQ; Lambert, Morton, et al., 2004). This 45-item self-report scale was designed to be administered during the course of therapy on a weekly basis as well as at termination. The OQ is comprised of three subscales: Symptom Distress (SD; measures symptoms of psychological disturbance, mainly depression and anxiety related), Interpersonal Relations (IR; measures satisfaction and problems in interpersonal relationships), and Social Role (SR; measures the patient's level of social role performance, such as problems at work). The OQ provides the total score, based on all 45 items and the subscale scores. The total OQ score represents the current level of distress of a given person, high scores indicating high level of distress. The OQ total scores were the only measure of treatment outcome in the aforementioned major six studies (brief explanation of these studies are provided below). Thus, stated in another way, the treatment outcome was operationally defined in the previous major studies as the level of distress measured by the total OQ scores.

Normative data of the OQ were obtained from several samples across various locations in

the United States. The details of the normative data are provided in the OQ manual (Lambert, Morton, et al., 2004). The total score has adequate three-week test-retest reliability (r = .84) and high internal consistency reliability (Cronbach's alpha = .93). The OQ has been shown to have strong concurrent validity, with validity coefficients ranging from .55 to .88 on SCL-90R, Beck Depression Inventory, Zung Self Rating Depression Scale, Zung Self Rating Anxiety Scale, Taylor Manifest Anxiety Scale, State Trait Anxiety Inventory, Inventory of Interpersonal Problems, Social Adjustment Scale, and the SF 36 Medical Outcome Questionnaire (Lambert, Morton, et al., 2004; Umphress, Lambert, Smart, & Barlow, 1997). The OQ has also been found to be sensitive to change over short periods of time when taken by patients in treatment (Vermeersch, Lambert, & Burlingame, 2000; Vermeersch et al., 2004) while untreated individuals' scores remained stable (Durham, 1999), thus demonstrating that this instrument is suitable for tracking change in patient disturbance.

Defining negative outcome and positive outcome. One of the key features of patient-focused methodologies involves the establishment of cutoff scores on outcome measures to determine if a given patient's change should be considered reliable or clinically significant. Using the formulas developed by Jacobson and Truax (1991), the OQ scores of both clinical and non-clinical samples were analyzed to provide cutoff scores for a Reliable Change Index (RCI) for the total and subscale scores (Lambert et al. 2004). The RCI of the total OQ score was 14. This means that a patient who makes a change equal to or greater than this value is considered to have reliably improved or reliably deteriorated, depending on the direction of the change. In addition to calculating the RCI value for the clinical significance cutoff score for the OQ total scores, a score demarcating the clinical population from the non-clinical population was calculated to be 63/64. This means that a person with an OQ total score of 63 or less is more

likely to come from the non-clinical population whereas an individual with a total OQ score of 64 or above indicates that the individual is more likely from the clinical population. These two criteria (RCI and clinical significance cutoff score of the OQ total scores) are used to determine negative and positive outcomes.

Defining identification of potential treatment failures. Various methods have been tried to determine the best way to accurately identify treatment failures (Lambert et al., 2003). Two methods (rational and empirical) have been demonstrated to be particularly effective in predicting the final treatment outcomes (OQ total scores at termination). The *rational* method is based on the algorithms that use information regarding the patient's early response to treatment, the dose response relationship, and the reliability of the OQ-45 (Lambert, Morton, et al., 2004; The details of the rational methods are provided in Lambert, Whipple, Bishop, et al., 2002). The *empirical* method utilized expected recovery curves for making predictions about final treatment outcomes. Standard recovery curves were developed based on the OQ total scores of 11,492 individuals with two or more OQ administrations from various clinical settings across the US (see Finch, Lambert, & Schaalje 2001). Based on the severity of distress, the full range of OQ total scores (0 to 180) were divided into 50 distinct groups.

These groups were created by first rank ordering all of the initial OQ total scores. Hierarchical linear modeling techniques were used to estimate the recovery curve for each of the OQ score groups. In addition to the calculation of the estimated recovery curves, tolerance intervals were calculated to allow for the identification of OQ total scores that are outside of the upper and lower limits of tolerance intervals for each given session. Sets of two tolerance intervals were established for the mean OQ scores at each session to identify unexpected change in progress in both positive and negative directions. These two-tailed intervals were set at 68%

and 80%. These intervals provided cutoff scores at each session for identifying 16% and 10% of the patients who were likely to fail in therapy or drop out prematurely. The details of the establishment of the expected recovery curves are described in the article by Finch, Lambert, & Schaalje (2001). As already mentioned above, the accuracy and clinical utility of the OQ algorithms for identifying at risk individuals for treatment failure have been demonstrated by Hannan and colleagues (2005), Lambert, Whipple, Bishop, et al. (2002), and Spielmans, Lambert and Masters (2006).

Patient progress feedback. As Lambert, Harmon, Slade, Whipple, and Hawkins (2005) stated, "the essence of improving outcomes for poorly responding patients is the development of a signal-alarm system that attempts to identify the failing patient before termination occurs" (p. 168). In the OQ feedback system, patients are administered the OQ at each session, beginning at the intake until termination. Based on the patients' session-by-session OQ data, feedback regarding each patient's predicted functioning at termination was provided to the clinicians in a form of a progress graph with the patient's past OQ scores up to the most recent OQ administration, color-coded message (green, white, yellow, and red) to catch the clinician's attention and to quickly convey the patient's progress status, and the feedback message corresponding to the color-coded message (The details of the messages are found elsewhere; e.g., Lambert et al., 2003; Lambert et al., 2005). The form of feedback has evolved as methodological improvements were made to the feedback studies. The first five feedback studies (studies 1, 2, 3, 4, and 5) used this form of feedback to clinicians. Most recently, the entire feedback system has been computerized by the use of the "OQ-Analyst" to allow for generation of immediate feedback at the beginning of each therapy session as patients fill out the OQ before meeting with their therapist. The feedback given to the clinicians essentially includes

all of the elements of feedback from the earlier methodology with the addition of the patient's responses on several critical items (e.g., suicide potentiality item), the patient's subscale scores, and the patient's current OQ total score in relation to various norm scores (i.e., community norm, outpatient norm, and inpatient norm).

Summary of Past Developments

To summarize the past six feedback studies, several acronyms are used to identify the experimental conditions. As patients entered treatment and subsequently participated in the past studies, they were assigned to one of the following conditions (see Figure 1 for detailed representation of all conditions used in the past studies): therapists of patients received OQ-based feedback (Fb); both therapists and patients received feedback (P/T Fb); and therapists of patients did not receive feedback or TAU (No-Fb). As treatment continued, patients divided themselves into two groups based on their treatment progress as measured by the OQ. Patients *not* progressing as expected (signal alarm cases; either red or yellow signal) were classified as "not-on-track" (NOT). Patients who progressed as expected (i.e., white or green signals) were termed "on-track" (OT). In recent studies, of those patients who were in the Fb and T/P Fb conditions *and* later identified as NOT, half the patients were given an additional intervention—"Clinical Support Tools" or *CST*. Accordingly, these groups of patients were termed *NOT-Fb+CST* or *NOT-T/P Fb+CST*.

In the first in a series of similar studies, Lambert, Whipple, et al. (2001) studied the effects of delivering feedback to therapists in a university counseling setting. This study included 609 participants, randomly assigned either to the experimental (Fb) condition (n = 307) or to the control (No-Fb) condition (n = 302). A main effect was found of keeping patients in

treatment longer, as well as improving outcome of the participants in the Fb condition relative to the clients in the control group (No-Fb) who were predicted to be treatment failures (NOT).

A replication study (study 2; Lambert, Whipple, Vermeersch, et al., 2002) was conducted on 1020 participants in the same university counseling center setting. In this study, participants were assigned to the experimental (Fb) group or the control (No-Fb) group depending on the semester in which they sought services. Thus, random assignment was not used. However, the authors reported that the groups were demographically equivalent. This study used 49 therapists (22 doctoral level psychologists and 27 doctoral students in training, including interns). Therapist assignment effect was demonstrated to be controlled by having approximately the same number of clients in both experimental and control groups and all therapists participating in both conditions. The results showed increased duration of treatment and improved outcome for NOT cases, thus replicating the findings of study 1. Nearly twice as many clients in the Fb condition reached clinically significant or reliable improvement and fewer actually deteriorated at termination.

In study 3, Whipple et al. (2003) hypothesized that strengthening the effects of feedback might improve the treatment outcome as well as the number of sessions attended by the individuals who were predicted to be treatment failures. To bolster the feedback effects, the researchers introduced the notion of Clinical Support Tools (CST) as a problem solving strategy to inform the therapists of factors that may be obstructing treatment. Their CST consisted of self-report measures of therapeutic alliance, client perceived social support, and readiness to change, and a decision tree to help clinicians examine these factors in a hierarchical fashion. The decision tree also advised for possible diagnostic reformulation and/or medication referral to a medical professional as appropriate. Whipple et al. (2003) found that clients in the CST

condition, who were identified as signal cases (NOT-Fb+CST), had better outcome than the signal cases whose clinicians either received feedback on their treatment response (NOT-Fb) or did not receive feedback (NOT-No-Fb). However, a methodological limitation was the failure to randomly assign participants to the CST experimental conditions.

In study 4, Hawkins et al. (2004) applied OQ feedback in a hospital-based outpatient setting, further hypothesizing that providing outcome feedback to patients themselves in addition to clinicians may have incremental value in enhancing outcome. Participants were 201 patients and five therapists (out of 715 patients invited to participate, 313 consented, 112 were excluded for not meeting the inclusion criteria of having at least two therapy sessions or by therapists' discretion for the possible iatrogenic effects of giving feedback to patients in four cases). Participants were randomly assigned to various treatment groups, using a randomized block design, with therapists serving as the blocking variable. The researchers found an effect for the clinician feedback condition as well as the clinician and patient feedback condition, and that the clinician and patient feedback condition demonstrated an incremental treatment effect. It should be noted that the NOT-Fb condition in this study did not include a CST feedback condition.

In study 5, Harmon et al. (2007) conducted a replication study, combining the experimental conditions used in studies 3 and 4 in a counseling center setting. This study used more than 1,374 participants. This study included random assignment to the CST condition, thus making an improvement over the Whipple methodology (study 3). The findings replicated the feedback effects on improved outcome in NOT groups. However, the effects of giving feedback to clients did not replicate the findings of study 4 (Hawkins, et al., 2004). In this study, the outcome of the patients in the Fb condition and that of the T/P Fb condition did not show a statistically significant difference in both OT and NOT groups. One methodological limitation

should be mentioned regarding this and the subsequent study. Based on the consistent findings indicating the benefits of providing OQ-based feedback to clinicians, the counseling center where the series of studies were conducted adopted routine provision of OQ-45 feedback to clinicians as part of their standard of care at the same time the Harmon study (study 5) commenced. This change in the center's policy prevented implementation of the treatment as usual (TAU) condition (i.e., No-Fb condition) in this and the subsequent study, thus making direct comparisons between the No-Fb condition and various experimental conditions no longer available.

In the most recent study (study 6), Slade et al., (2008) made improvements in the CST measures and added computer-generated real-time feedback via the OQ-Analyst. Accordingly, this study addressed the question of whether immediate, computer-generated feedback is superior to time-delayed feedback in improving outcome for NOT clients. This study also employed an alternative self-report scale for measuring client motivation for therapy and incorporated an additional self-report scale of client's level of perfectionism. Furthermore, this study incorporated the therapist-patient feedback condition as Studies 4 and 5 did. Results from the Slade study suggest that the use of the revised CST improved outcome, again replicating the work of Harmon and Whipple, but failed to find an effect for combined therapist and client progress feedback.

After the completion of study 3, Lambert and his colleagues conducted a small meta-analysis of the first three studies (Lambert et al., 2003). In this study, the researchers found that overall the Fb conditions had a deterioration rate of 5% whereas the No-Fb condition resulted in 9% deterioration rate. They also found that, although the overall improvement rates of the feedback and no-feedback conditions did not differ greatly, of those who were identified as NOT

cases, the No-Fb group had a 21% deterioration rate while the Fb group showed a deterioration rate of 13%. Furthermore, of the individuals identified as NOT cases, those in the Fb condition had greater reliable and clinically significant change rates (35%) than the signal cases in the TAU condition (21%). These results suggested substantial beneficial effects of progress feedback especially on those clients who were at risk of failing treatment.

Now that three more major studies have been completed with further developments in the feedback system, and that the effects of the CST have not been summarized across studies, conducting another review appeared appropriate. Furthermore, the 2003 meta-analysis focused mainly on the evaluation of the effect sizes associated with Fb and No-Fb conditions thereby essentially addressing questions regarding the amount of effects without addressing questions about the patterns of change in patient outcome across sessions, such as the rate of recovery. The overall purpose of the current study was to conduct a quantitative summary of the past studies on the OQ-45 based quality assurance system. Before describing the research questions, however, a few notes on the methodology of this study should be made.

Methodological Considerations

Meta-analysis has become a popular method of reviewing research results over the last two decades in various scientific disciplines (Rosenthal & DiMatteo, 2001). Although meta-analytic methods have become more sophisticated over the years, there is no consensus on a single method (Rosenthal & DiMatteo, 2001). Meta-analysis is a quantitative method for combining numerical results from multiple studies and allows researchers to arrive at more precise and replicable conclusions than those derived from any single study or a non-quantitative, narrative account (Rosenthal & DiMatteo, 2001). The scope of the present study is limited to the quality assurance system developed by Lambert and his colleagues. At least two

primary reasons exist for this limited scope. The primary reason is that, to our knowledge, no other published studies have incorporated the methodology used by Lambert's team for the same purpose, which is to enhance the treatment outcome of signal-alarm clients through the use of feedback interventions. Thus, broadening the scope of the present study (e.g., investigating the effects of feedback interventions in general) cannot address the questions of our interest. Another reason is that a recent meta-analysis of feedback interventions in health services, including studies that examined the effects of providing feedback to mental health professionals of their clients' status, has already been conducted. Sapyta, Riemer, and Bickman (2005), referring to their recent meta-analysis, reported the average effect size of feedback interventions was 0.21. When the effect of feedback on "flagged" clients (our signal-alarm cases) was compared to feedback on not flagged clients (our on-track cases), feedback interventions to flagged samples had an effect size of 0.31.

A meta-analysis takes either fixed- or random-effects models, depending on the nature of inferences the researcher wish to make (Hedges & Vevea, 1998). Fixed models are appropriate in the following circumstances:

> If the analyst wishes to make inferences only about the effect-size parameters in the set of studies that are observed (or to a set of studies identical to the observed studies except for uncertainty associated with the sampling of participants into those studies) . . . [fixed]-effects analysis procedures are appropriate for making conditional inferences (Hedges & Vevea, 1998, p. 487).

Thus, as Hedges and Vevea (1998) further explained, meta-analysis based on fixed-models cannot make any inference about any other studies that "may be done later, could have been

done earlier, or may have already been done but are not included among the observed studies (p. 487)." Thus, the inferential ability of fixed-effects models is very limited.

Random-effects models are utilized when the researcher "wishes to make inferences about the parameters of a population of studies that is larger than the set of observed studies and that may not be strictly identical to them" (Hedges & Vevea, 1998). Regarding the use of random-effects models, Overton (1998) recommended the following:

> When the contextual conditions are ill defined (e.g., a relatively new research area or one in which previously ignored contextual factors are being discovered to be important), a mixed model is advised to account for the uncertainty in the critical contextual determinants of the effect under study. Or if the sample domain notably underrepresents the population domain, again a mixed model is preferred because of the uncertainty expressed by the relatively limited sampling of conditions in the meta-analysis studies (p. 376).

Given that research on providing feedback for the purpose of enhancing treatment outcome is relatively new and under-explored in psychotherapy literatures and generalization is desirable, a random-effects model was employed in the present study.

The notion of non-independence of effects (Rosenthal & DiMatteo, 2001) should also be addressed in the present study because of the homogeneous characteristics of the studies included in the present analysis (i.e., Five of the six studies were conducted at the same university counseling center setting, conducted by the same research team, and all studies were conducted by the same research team.).

Study 1: Meta- and Mega-Analyses of OQ Total Scale Score-Based Outcome

Now that three more major studies have been completed after the first small meta-analysis by Lambert et al. (2003a) with further developments in the feedback system, and that the effects of the P/T Fb and CST Fb interventions have not been summarized across studies, conducting another research synthesis study appeared appropriate. As repeatedly demonstrated in the previous feedback studies, OQ feedback interventions appear to be effective in enhancing outcome for NOT patients, while having little impact on OT cases. Thus, the primary purpose of this meta-analysis was to investigate the effects of various OQ feedback interventions on patient outcome whose progress was identified as NOT. Although subtle differences existed in the operationalization of feedback interventions across studies, given similarities in methodologies, all of the feedback interventions were grouped in one of the following:

- NOT Feedback (NOT Fb):[1] NOT patients whose OQ progress feedback was provided to their therapists only.
- NOT Patient/Therapist Feedback (NOT P/T Fb):[2] NOT patients whose OQ progress feedback was provided directly to both patients and therapists.
- CST Feedback (CST Fb):[3] NOT patients whose OQ progress feedback and Clinical Support Tools were provided to their therapists.
- NOT Treatment As Usual (TAU):[4] NOT patients whose therapists received no feedback intervention.

[1] Studies 1, 2, 3, and 4 included the NOT Fb condition.
[2] Studies 4, 5, and 6 included NOT P/T Fb condition.
[3] Studies 3, 5, and 6 utilized Clinical Support Tools Feedback interventions. Due to study designs, studies 5 and 6 employed variations of CST Feedback groups: CST Feedback group, Patient/Therapist CST Feedback group, a week delayed CST Feedback group, and two weeks delayed CST Feedback group. Due to statistically non-significant findings between the CST groups, we combined them as CST Feedback group in this meta-analysis.
[4] Studies 1, 2, 3, and 4 included the NOT TAU condition.

- OT Feedback (OT Fb):[5] OT patients whose OQ progress feedback was provided to their therapists only.
- OT Patient/Therapist Feedback (OT P/T Fb):[6] OT patients whose OQ progress feedback was provided directly to both patients and therapists.
- OT Treatment As Usual (OT TAU):[7] OT patients whose therapists received no feedback intervention.

Selection Criteria and Participants

All of the six OQ feedback studies published to date were included in this analysis. Each study's demographic variables, mean OQ total score at pre-treatment, and *n* and percentage of patients identified as NOT cases are reported in Table 1 in Appendix A (All tables are presented in Appendix A). Of the 6,151 cases included in this study, 355 patients were included multiple times either in different studies or within the same study. Inclusion of data from multiple episodes of care for the same individuals violates the statistical assumption of independence of observation. In routine clinical practice, however, patients frequently return to treatments for various reasons. Thus, inclusion of all data was deemed to yield results that are more representative of naturalistic treatment settings. Analyses were conducted both with and without the patients who appeared multiple times. Because results were nearly identical, only the results from analyses with patients of multiple appearances are reported unless indicated otherwise.

The statistical methods used in the previous feedback studies reflect two distinct approaches: effectiveness analysis based on the intent-to-treat (ITT) principle and efficacy analysis (Lachin, 2000; Atkins, 2009). These approaches reflect two distinct philosophies in terms of the interpretation of their results. The former addresses the overall effect of a treatment

[5] All six studies included the OT Fb condition.
[6] Studies 4, 5, and 6 included the OT P/T Fb condition.
[7] Studies 1, 2, 3, and 4 included the OT TAU condition.

at the population level, regardless of various treatment compliance issues that may arise in naturalistic clinical settings. This method essentially includes the data of all patients solely on the basis of the initial assignment to treatment conditions. The latter approach addresses the effect of a given treatment on a subset of patients who met certain compliance criteria to be considered "completers" of the treatment regimen. *The studies, which examined the effects of the Fb intervention against the TAU employed the effectiveness analyses. Alternatively, two of the three studies that employed the CST Fb condition applied post-hoc screening criteria to analyze a subset of patients who completed the prescribed feedback interventions. Given these differences in analytical approaches, each feedback treatment under both approaches was evaluated using the original datasets of all six studies included in this study.*

In the ITT analyses, all participants in the CST Fb, NOT P/T Fb, NOT Fb, NOT TAU, OT P/T Fb, OT Fb, and OT TAU groups were included. These analyses provide the most conservative estimates of the treatment effects, as they even incorporate the data of individuals whose post-treatment scores are missing, including the data of those with only the intake and warning OQ scores. Patients with only one data point were grouped within the OT groups. To obtain conservative estimates of these patients' post-test scores, their last observed data point (or their only data point) was carried forward and treated as their post test score, utilizing the last observation carried forward method. The breakdown of the number of participants in each treatment condition across all six studies was as follows: NOT Fb (n = 427), NOT P/T Fb (n = 222), CST Fb (n = 415), NOT TAU (n = 318), OT Fb (n = 2390), OT P/T Fb (n = 935), and OT TAU (n = 1445).

In the efficacy analyses, the inclusion criteria were retrospectively defined to represent the least necessary condition in which the effects of the OQ feedback interventions (i.e., Fb, P/T

Fb, and CST Fb) could be measured.[8] For the analyses of CST Fb interventions, the exclusion criteria as defined in the original articles in studies 5 and 6 (Harmon et al., 2007; Slade et al., 2008) were used. Study 3 was the first study to implement the CST intervention; however, this study did not employ exclusion criteria similar to those applied in studies 5 and 6. Thus, the minimum inclusion criteria[9] required for a given patient to be considered a completer of the CST intervention in study 3 were retrospectively defined and applied. Through the application of the aforementioned inclusion criteria, the following number and percentage of participants were included in each treatment condition when aggregated across studies: NOT Fb, n = 263 (61.6%); NOT P/T Fb, n = 177 (79.7%); CST Fb, n = 217 (52.2%); OT Fb, n = 1651 (69.0%); OT P/T Fb, n = 777 (83.1%).

Dependent Measures and Computation of Effect Sizes

The effects of OQ feedback interventions were compared on the following dependent measures: Mean post-treatment OQ total score, the odds of patients achieving clinically significant improvement at post-treatment, the odds of the occurrence of clinically significant worsening (or deterioration) at post-treatment, mean pre-post change scores, and between group differences in pre-post change scores. Mean number of sessions attended by patients in each

[8] The efficacy sample inclusion criteria for the NOT Fb and NOT P/T Fb groups were defined as follows: attended at least five sessions (for studies 1, 2, 3, 4, and 5) or four sessions (for study 6 due to electronic immediate progress feedback), completed the OQ in at least three sessions, and the last recorded OQ score came from at least two sessions (for studies 1, 2, 3, 4, and 5) or one session (for study 6) after the patient was identified as a NOT case. The efficacy sample inclusion criteria for the OT Fb and OT P/T Fb groups were set more loosely than their NOT counterparts, given that a majority of OT patients left treatment before the effects of feedback treatments could be measured (i.e., nearly 70% attended four or less sessions). Accordingly, the OT Fb and P/T Fb criteria were defined as the following: attended at least two sessions and filled out the OQ in at least two of the sessions attended.

[9] To identify a given patient as a NOT case, administer the CST intervention, and measure the effects of the CST intervention in study 3, the patient needed to have attended at least six sessions (three of which occurred after the patient was identified as a NOT case) and filled out the OQ in at least three of the sessions. Application of these inclusion criteria, nonetheless, does not guarantee the inclusion of only those who completed the CST intervention. For instance, NOT patients who attended more than required number of sessions, but failed to complete the OQ after the administration of the CST intervention would still be included in the analysis despite lacking the post-treatment score.

condition was also compared for the ITT analyses, but not for the efficacy analyses, as different numbers of sessions attended by patients were part of the inclusion criteria.

Following recommendations by Overton (1998) and Hedges and Vevea (1998), given that the research on providing feedback for the purpose of enhancing treatment outcome is relatively new in psychotherapy outcome literature; that studies included in this meta-analysis contained slight variations in research designs; and that the purpose of the present study was to investigate the applicability of our findings to a broader clinical contexts, a random-effects model was employed. Hedges's standardized mean difference *g* (Hedges, 1981) was used as the unit of effect size for mean post-treatment OQ total score comparisons and mean number of sessions attended by patients between feedback groups and control groups. Formulae for obtaining Hedges's *g* are provided in Appendix B. Random weights were then assigned to individual standardized mean differences to obtain the estimated weighted mean effect size per comparison. Formulae for calculating random weights and estimated weighted mean effect sizes (or combined effect sizes) are presented in Appendix C. As lower OQ scores indicate lower levels of distress, negative effect sizes in post-treatment OQ total scores comparisons signify superior outcome of the treatment condition in question. Although the P/T Fb and CST Fb groups were not directly compared against the TAU condition in some of the studies, such comparisons were considered to provide more intuitive interpretation regarding the effects of feedback interventions in relation to the TAU condition. Thus, mega-analyses on the pooled dataset from all of the six feedback studies were conducted to calculate the effect sizes of feedback interventions (i.e., P/T Fb and CST Fb) in relation to TAU. Such an approach with large *n* provides an alternative method to traditional meta-analysis in research synthesis (e.g., DeRubeis, Gelfand, Tang, & Simons, 1999; Serretti, Cusin, Rausch, Bondy, & Smeraldi, 2006).

Possible heterogeneity of effect sizes and publication biases were tested. Given the small number of studies included in this study, mega-analytic approaches were used to test for the homogeneity of effect sizes. To test for heterogeneity of effect sizes in mean post-treatment OQ score differences, separate analyses of covariance were performed for each pooled treatment group, with study as the factor, post-treatment OQ total score as the dependent variable, and pre-treatment OQ total score as the covariate. To test for equivalence in pre-treatment distress level across groups, independent samples *t*-test for each between-group comparison was conducted. To test for heterogeneity of effect sizes in mean number of session attendance, one-way analyses of variance for each pooled treatment group were conducted, with number of session attendance as the dependent variable and study as the factor. Classic fail-safe *N* test (Rosenthal, 1979), Orwin's (1983) fail-safe *N* test, and Duval and Tweedie's (2000) trim and fill were performed to address possible publication biases.

Another set of treatment outcomes investigated in this study was differences in proportions and odds of patient outcome classification based on clinical significance indices. The use of clinical significance indices based on a clinical cutoff score and reliable change index methods proposed by Jacobson and Truax (1991) is one of the hallmarks of the OQ-45-based quality assurance system. As demonstrated by Beckstead et al. (2003) and Lunnen and Ogles (1998), the OQ-45-based clinical significance classification of patient outcome appears to reflect meaningful change as well as the functional/dysfunctional state of patients. In this quantitative review, the clinical significance status for each patient at termination was classified in one of the three categories: deterioration/reliable worsening, no change, or clinically significant improvement. When examining the differences in the proportion of patients who experienced deterioration among two groups in comparison (e.g., feedback versus TAU), patients in each

group were first dichotomized into either deterioration cases or non-deterioration cases. Odds of patients experiencing deterioration were calculated for each group in comparison. The odds of deterioration among the treatment group were divided by the odds of deterioration among the control group to yield the odds ratio (*OR*). When the odds of deterioration are identical for both groups, the resulting *OR* is 1. If the odds of deterioration among the treatment group are lower than those in the control group, the resulting OR is less than 1. Accordingly, if the odds of deterioration among the treatment group are higher than the control group, the resulting *OR* is higher than 1. To examine the difference in the odds of deterioration, statistical significant test were conducted with the alpha of .05. Similar procedures were followed when the odds of clinically significant improvement were examined.

The results are presented in three ways. First, *n* and percentage of patients in each of the three clinical significance categories for each feedback intervention group across all six studies were aggregated and reported. Second, the odds of the occurrence of deteriorated/reliably worsened cases were compared for each feedback intervention group against its control group (i.e., TAU or Fb groups, depending on the comparisons being made) in the unit of odds ratio. Third, the odds of the occurrences of clinically significant improvement were similarly compared for each feedback intervention against its control. To expedite the statistical calculations, Comprehensive Meta-Analysis, Version 2 was utilized in the calculation of effect sizes.

Analyses of Effect Sizes

Effects of Feedback Interventions on Post-treatment OQ Total Score in NOT Patients. The combined effect size and the results of tests of publication bias for each of the comparisons presented below are summarized in Table 2.

Feedback (Fb) effect. The results of one-way ANCOVA, testing for heterogeneity of effects across studies, with study as the factor, post-treatment OQ total score as the dependent variable, and pre-treatment OQ total score as the covariate, indicated no significant study effect among the Fb group in ITT analysis, $F(5, 420) = 0.221$, $p = 0.951$, or efficacy analysis, $F(5, 250) = 1.49$, $p = 0.192$. However, statistically significant study effect was found for the TAU group, $F(3, 313) = 2.79$, $p = 0.041$. The result of the independent samples *t*-test of pre-treatment mean OQ scores between pooled Fb and TAU groups was not significant in ITT analysis, $t(743) = -.28$, $p = .778$, or efficacy analysis, $t(579) = -0.48$, $p = 0.631$, indicating Fb and TAU were comparable at pre-test distress level. Thus, despite the heterogeneity among the TAU groups in mean post-test scores, given equivalent pre-treatment OQ scores across groups, proceeding with aggregating the TAU data was deemed appropriate in favor of ecological validity. ITT meta-analysis indicated that effect sizes of individual studies comparing the NOT Fb and NOT TAU groups ranged from $g = -0.42$, $p < 0.001$, 95% CI [-0.68, -0.17] to $g = 0.08$, $p = 0.742$, 95% CI [-0.41, 0.58] (See Table 1 in Appendix D for complete list of individual effect sizes and the forest plot). The aggregate effect size was statistically significant at the .05 level, $g = -0.28$, $p = 0.003$, 95% CI [-0.47, -0.10]—equivalent of 6.4 OQ total score difference on average. When the efficacy sample inclusion criteria were applied to the same comparison groups, the results showed greater treatment effect favoring the Fb intervention. As shown in Table 2 in Appendix D, effect sizes for individual studies ranged from $g = -0.78$, $p < 0.001$, 95% CI [-1.11, -0.45] to $g = -0.18$, $p = 0.523$, 95% CI [-0.73, 0.37]. The aggregate effect size was significant at the .05 level, $g = -0.53$, $p < 0.001$, 95% CI [-0.78, -0.28], which equates to 12.0 OQ total points difference on average.

Patient/therapist feedback (P/T Fb) effect. Although the ideal evaluation would have been to compare all of the feedback interventions against the TAU group, the last two feedback studies (studies 5 and 6) containing the P/T Fb groups did not have TAU groups, as explained earlier. Thus, P/T Fb groups were compared against Fb groups, where Fb groups were used as the benchmark to evaluate incremental benefits of the P/T Fb intervention. The results of one-way ANCOVA to test the heterogeneity of effects showed that study effect did not reach statistical significance for the P/T Fb groups in ITT analysis, $F(2, 218) = 1.58$, $p = 0.208$, or efficacy analysis, $F(2, 173) = 1.62$, $p = 0.201$. As presented in Table 3 in Appendix D, ITT analyses of NOT P/T Fb vs. NOT Fb indicated none of the individual effect sizes were significant at the .05 level, with individual effect sizes ranging from $g = -0.44$, $p = 0.071$, 95% CI [-0.92, 0.04] to $g = -0.10$, $p = 0.526$, 95% CI [-0.39, 0.20]. The aggregate effect size also did not reach statistical significance at the .05 level, $g = -0.16$, $p = 0.099$, 95% CI [-0.36, 0.03]. As presented in Table 4 in Appendix D, when the efficacy criteria were applied, individual effect sizes ranged from $g = -0.39$, $p = 0.177$, 95% CI [-0.96, 0.18] to $g = -0.06$, $p = 0.734$, 95% CI [-0.40, 0.28]. The aggregated effect size was similar to that of the ITT analysis, $g = 0.16$, $p = 0.163$, 95% CI [-0.37, 0.06]. These results suggest that, in terms of treatment outcome at termination, providing progress feedback to both clinicians and patients adds no significant incremental benefit to providing progress feedback only to clinicians (who may or may not share it with patients).

Pre-treatment mean OQ total score comparison between the pooled P/T Fb and pooled TAU groups did not reach statistical significance for either ITT analysis, $t(538) = 0.65$, $p = 0.518$, or efficacy analysis, $t(493) = .24$, $p = 0.810$, indicating that the two groups did not differ significantly in their initial level of disturbance. ITT post-treatment score difference was

significant at the .05 level, g = -0.36, p <0.001, 95% CI [-0.54, -0.19], equivalent of 7.9 points difference in mean OQ total scores. Efficacy post-treatment score difference was also significant, g = -.55, p <0.001, 95% CI [-0.73, -0.36], equivalent of mean OQ total score difference of 11.7 points. These results suggest that NOT patients in the P/T Fb condition experience greater therapeutic gain as measured by the OQ-45 at termination than those in TAU. Such therapeutic benefits are more pronounced among those who stayed in treatment long enough to experience the benefits of P/T Fb intervention.

Clinical support tools feedback (CST Fb) effect. As in the case of P/T Fb analyses, two of the three studies that tested the effects of the CST interventions (studies 5 and 6) did not employ the TAU condition. Thus, the CST Fb groups were also compared to the Fb groups to estimate their incremental clinical benefits over the Fb condition. The results of one-way ANCOVAs to test for heterogeneity of effects for the CST Fb group did not reach statistical significance in ITT analysis, $F(2, 411) = 2.00$, $p = 0.137$ or efficacy analysis, $F(2, 213) = 0.48$, $p = 0.617$. As presented in Table 5 in Appendix D, the ITT analysis indicated that individual effect sizes ranged from g = -0.23, p = 0.094, 95% CI [-0.49, 0.04] to g = -0.11, p = 0.415, 95% CI [-0.38, 0.16]. The combined effect size was significant at the .05 level, g = -0.16, p = 0.048, 95% CI [-0.33, -0.002], indicating approximately 3.6 OQ total points difference on average, favoring the CST Fb group. When the efficacy criteria were applied to both the CST Fb and Fb groups, the combined effect size was g = -0.19, p = 0.113, 95% CI [-0.43, 0.05], equivalent of approximately 4.2 OQ total points difference on average. Individual effect sizes ranged from g = -0.32, p = 0.053, 95% CI [-0.65, 0.01] to g = -0.11, p = 0.606, 95% CI [-0.32, 0.55] (Table 6 in Appendix D). It should be pointed out that, contrary to the outcome comparison between the CST Fb group and the Fb group reported in study 3 (Whipple, et al., 2003), which reported

results favoring the CST Fb group, application of the efficacy criteria in this study yielded a result favoring the Fb group. Although the two groups appeared demographically similar at pre-treatment, given that random assignment of NOT patients to CST Fb and Fb groups was not employed in this study, such contradictory findings may have been due to unknown artifacts resulting from therapists' selection of patients into treatment conditions. When study 3 was removed from the efficacy analysis, the aggregate effect size of the CST Fb group over Fb group improved to g = -0.29, p = 0.013, 95% CI [-0.52, -0.06], equivalent of approximately 6.2 OQ point difference on average. These results suggest that, on average, those NOT patients who receive the CST intervention in routine care in addition to the Fb intervention experience small additional therapeutic gains represented in about 3 to 4 OQ points reduction over those who receive only progress feedback intervention alone. Those who stay in treatment long enough to experience the benefit of the CST intervention, on average, experience further distress reduction over those who experience the benefit of the Fb intervention alone. More studies employing random assignment-based comparison between the CST Fb and Fb conditions may help us better estimate the effect of the CST Fb intervention.

Pre-treatment mean OQ total score comparison between the pooled CST Fb and pooled TAU groups did not reach statistical significance for either ITT analysis, $t(731) = -0.34$, p = 0.732, or efficacy analysis, $t(533) = 0.73$, p = 0.468, indicating the two groups were comparable at pre-treatment. ITT analysis indicated that post-treatment score difference between CST Fb group and NOT TAU was significant at the .05 level, g = -0.44, $p < 0.001$, 95% CI [-0.59, -0.30], equivalent of 9.5 points difference in mean OQ total scores. Efficacy post-treatment score difference was also significant, g = -.70, $p < 0.001$, 95% CI [-0.88, -0.52], equivalent of mean OQ total score difference of 14.6 points. These results suggest NOT patients

who receive the CST Fb intervention, on average, experience significantly more therapeutic gain than those in the TAU condition. Such therapeutic gain is more pronounced among who stay in treatment long enough to see the benefit of the CST Fb intervention.

Effects of feedback interventions on clinical significance. The *n* and percentage of the clinical significance classification of patient outcome at termination were aggregated by each treatment condition and are presented in Table 3. The summary of combined effects for the odds of deterioration/reliable worsening and the results of tests of publication bias are presented in Table 4. The summary of combined effects for the odds of clinically significant improvement and the results of tests of publication bias are presented in Table 5.

Feedback (Fb) effect. When the odds of patient deterioration/reliable worsening at termination of the NOT Fb group were compared against NOT TAU, the results of ITT analyses indicated that the combined effect was significant at the .05 level, *OR* = 0.62, *p* = 0.040, 95% CI [0.40, 0.98], with effect sizes of individual studies ranging from *OR* = 0.21, *p* = 0.063, 95% CI [0.04, 1.09] to *OR* = 0.72, *p* = 0.315, 95% CI [0.39, 1.32] (Table 7 in Appendix D). When the efficacy criteria were applied to the Fb group, the combined odds of deterioration for the Fb group decreased to *OR* = 0.44, *p* = 0.015, 95% CI [0.23, 0.85], with odds ratios of individual studies ranging from *OR* = 0.21, *p* = 0.041, 95% CI [0.05, 0.94] to *OR* = 0.60, *p* = 0.238, 95% CI [0.25, 1.41] (Table 8 in Appendix D). These results suggest that the odds of deterioration among NOT patients in TAU are approximately 1.5 times higher than those who received the Fb intervention in routine practice. The results further suggest that the odds of deterioration among TAU are about 2.3 times higher than those who had stayed in treatment long enough to receive the benefit of the Fb intervention. When the odds of patients achieving clinically significant improvement at termination were compared between the NOT Fb group and NOT TAU, the

results indicated a significantly increased odds at the .05 level favoring the Fb group, *OR* = 1.70, *p* = 0.005, 95% CI [1.17, 2.46], with individual effect sizes ranging from *OR* = 1.44, *p* = 0.539, 95% CI [0.45, 4.65] to *OR* = 2.17, *p* = 0.012, 95% CI [1.19, 3.97] (Table 9 in Appendix D). When the efficacy criteria were applied, the combined odds ratio of the occurrence of clinically significant improvement among the NOT Fb group against the NOT TAU group was *OR* = 2.55, *p* < 0.001, 95% CI [1.64, 3.98], with odds ratios of individual studies ranging from *OR* = 1.23, *p* = 0.766, 95% CI [0.32, 4.67] to *OR* = 2.97, *p* = 0.003, 95% CI [1.44, 6,11] (Table 10 in Appendix D). These results suggest clinical benefit of the Fb intervention in reducing the occurrence of treatment failure while increasing the odds of patients experiencing clinically significant improvement.

Patient/therapist feedback (P/T Fb) effect. When the odds of the occurrence of deterioration/reliable worsening were compared between the NOT P/T Fb and NOT Fb groups, the results of ITT analyses indicated that effect sizes of individual studies ranged from *OR* = 1.00, *p* = 1.000, 95% CI [0.42, 2.38] to *OR* = 1.74, *p* = 0.184, 95% CI [0.77, 3.95] (Table 11 in Appendix D). The combined effect size was not statistically significant, *OR* = 1.35, *p* = 0.306, 95% CI [0.758, 2.413]. When the efficacy criteria were applied, the combined odds ratio of deterioration cases increased for the P/T Fb condition, although the results did not reach the .05 significance level, *OR* = 1.89, *p* = 0.094, 95% CI [0.90, 3.96]. Individual effect sizes ranged from *OR* = 0.68, *p* = 0.788, 95% CI [0.04, 11.53] to *OR* = 2.95, *p* = 0.047, 95% CI [1.02, 8.54] (Table 12 in Appendix D). Although statistical significance was not achieved, the results suggest a higher rate of deterioration among NOT patients in the P/T Fb condition than those in the Fb condition.

ITT comparisons of the odds of clinically significant improvement yielded the combined effect of *OR* = 1.44, *p* = 0.086, 95% CI [0.95, 2.19], with individual effect sizes ranging from *OR* = 1.26, *p* = 0.495, 95% CI [0.65, 2.47] to *OR* = 1.94, *p* = 0.179, 95% CI [0.74, 5.10] (Table 13 in Appendix D). The efficacy analyses indicated that the combined effect size was similar to that obtained from the ITT sample, *OR* = 1.38, *p* = 0.164, 95% CI [0.88, 2.18] with a similar range of individual study effect sizes, *OR* = 1.25, *p* = 0.521, 95% CI [0.63, 2.50] to *OR* = 1.56, *p* = 0.459, 95% CI [0.48, 5.00] (Table 14 in Appendix D). Although statistical significance was not reached, the results suggest higher odds of clinically significant improvement among NOT patients in the P/T Fb condition than those in the Fb condition. These results suggest that provision of direct progress feedback to NOT patients has potential clinical effects that may enhance outcome in some patients even beyond what can be achieved by provision of progress feedback to clinicians alone, while having possible iatrogenic effect in some.

The odds of deterioration/reliable worsening between the pooled NOT P/T Fb group and pooled NOT TAU group did not reach statistical significance in ITT analysis, *OR* = 0.74, *p* = 0.199, 95% CI [0.47, 1.17], or efficacy analysis, *OR* = 0.68, *p* = 0.134, 95% CI [0.42, 1.13]. The odds of clinically significant improvement between the pooled NOT P/T Fb group and pooled NOT TAU was significant in both ITT analysis, *OR* = 2.20, *p* < 0.001, 95% CI[1.51, 3.21] and efficacy analysis, *OR* = 2.87, *p* < 0.001, 95% CI [1.93, 4.27]. These results suggest that the P/T Fb intervention, in comparison to TAU, does not decrease the odds of deterioration, but increases the odds of improvement among NOT patients.

Clinical support tools feedback (CST Fb) effect. ITT Comparisons between the CST Fb groups against the NOT Fb groups indicated that individual effect sizes of deterioration/reliable worsening ranged from *OR* = 0.59, *p* = 0.342, 95% CI [0.20, 1.76] to *OR* = 1.043, *p* = 0.916,

95% CI [0.48, 2.29] (Table 15 in Appendix D) with the combined effect size of *OR* = 0.76, *p* = 0.288, 95% CI [0.46, 1.26]. The results of efficacy analyses indicated the combined effect was *OR* = 0.66, *p* = 0.329, 95% CI [0.29, 1.52] with individual effect sizes ranging from *OR* = 0.54, *p* = 0.356, 95% CI [0.15, 2.0] to *OR* = 0.83, *p* = 0.756, 95% CI [0.25, 2.74] (Table 16 in Appendix D). When the odds of patients achieving clinically significant improvement were compared in the ITT analyses, the combined effect size was *OR* = 1.53, *p* = 0.016, 95% CI [1.08, 2.18] favoring the CST Fb, with individual study effect sizes ranging from *OR* = 1.22, *p* = 0.467, 95% CI [0.69, 2.184] to *OR* = 1.97, *p* = 0.050, 95% CI [1.00, 3.87] (Table 17 in Appendix D). The results of the efficacy analyses of comparing the odds of patients achieving clinically significant improvement yielded the combined effect size of *OR* = 1.83, *p* = 0.098, 95% CI [0.89, 3.76], with individual effect sizes ranging from *OR* = 1.167, *p* = 0.729, 95% CI [0.487, 2.729] to *OR* = 3.610, *p* < 0.001, 95% CI [1.847, 7.057] (Table 18 in Appendix D). These results suggest that the CST Fb, in comparison to the NOT Fb increases the odds of patients achieving clinically significant improvement, but the odds of deterioration/reliable worsening does not seem to decrease, at least at a statistically significant level.

The odds of deterioration/reliable worsening between the pooled CST Fb group and the pooled NOT TAU group reached statistical significance in both ITT analysis, *OR* = 0.51, *p* = 0.001, 95% CI [0.34, 0.76] and efficacy analysis, *OR* = 0.23, *p* < 0.001, 95% CI [0.12, 0.44]. The odds of clinically significant improvement between the pooled CST Fb group and the pooled TAU was significant in both ITT analysis, *OR* = 2.20, *p* < 0.001, 95% CI [1.51, 3.21] and efficacy analysis, *OR* = 2.87, *p* < 0.001, 95% CI [1.93, 4.27]. The odds of clinically significant improvement between the same groups reached statistical significance in ITT analysis, *OR* = 2.01, *p* < 0.001, 95% CI [1.51, 2.92] and efficacy analysis, *OR* = 3.85, *p* < 0.001, 95% CI [2.65,

5.60]. These results indicate that the odds of patients in TAU experiencing deterioration are approximately 2.0 times higher than those receiving the CST Fb in routine care settings (ITT). When comparing against those who complete the CST intervention, the odds of deterioration/reliable worsening among the TAU patients is approximately 4.3 times higher than those in the CST group. The results further indicate that the odds of patients in the CST Fb group achieving clinically significant improvement in routine care settings (ITT) is approximately 2.0 times higher than those in TAU. The odds of clinically significant improvement among those who complete the CST Fb intervention are about 3.9 times higher than those in the TAU.

Pre-post change in OQ total scores. Prior to calculating combined pre-post effect sizes, series of one-way ANCOVAs were conducted to test the presence of heterogeneity among mean pre-post change scores across studies, using pre-post change score as the dependent variable, study as the factor, and pre-treatment OQ score as a covariate. The results of ANCOVA for NOT TAU was significant at .05 level, $F(3, 313) = 2.788$, $p = 0.041$. The cause of the heterogeneity among NOT TAU, which ranged from $M = -8.41$ ($SD = 18.32$) in study 4 to $M = 4.81$ ($SD = 21.23$) in study 1, is not known at this time. The results of ANCOVAs for NOT Fb was not significant in ITT analysis, $F(5, 420) = 0.223$, $p = 0.953$ or efficacy analysis, $F(5, 256) = 1.428$, $p = 0.214$. Similarly, heterogeneity was not observed among NOT P/T Fb group in either ITT analysis, $F(2, 218) = 1.583$, $p = 0.208$ or efficacy analysis, $F(2, 173) = 1.618$, $p = 0.201$. CST Fb groups also did not yield significant results in ITT analysis, $F(2, 411) = 1.995$, $p = 0.137$, or efficacy analysis, $F(2, 213) = 0.484$, $p = 0.617$. Given the heterogeneity among the NOT TAU groups, the consequences of the effect size calculation method (i.e., meta-analysis versus mega-analysis) were considered. To investigate such consequences of using one method

over the other, effect sizes were calculated and compared in both ways. The results showed that the smaller the between-study heterogeneity were (i.e., smaller *F* ratio found in the ANCOVAs reported earlier), the more similar the results between the mega-analysis and the meta-analysis. For instance, NOT Fb, which had the smallest *F* ratio of all treatment conditions, yielded the effect sizes and 95% CIs that were essentially identical. The results from the meta-analysis and mega-analysis were also similar in other treatment groups, although the results from the mega-analysis tended to yield somewhat smaller effect sizes and/or 95% CIs. Given the preference for generalizability of the results, the results from the meta-analysis seemed more appropriate. Thus, only the meta-analytic results are reported where both methods were utilized.

Pre-post change effect size was obtained for each treatment group and presented in Table 6. Effect sizes for between-group difference in the pre-post change effect size were then calculated for each between-group comparison of interest. In the ITT analysis, the pre-post effect size of the TAU was g = -0.04, p = 0.706, 95% CI [-0.23, 0.15]. The effect size for the NOT Fb group was g = -0.25, p < 0.001, 95% CI [-0.33, -0.16]. The effect size for NOT P/T Fb was g = -0.49, p < 0.001, 95% CI [-0.75, -0.22]. The effect size for the CST Fb group was g = -0.45, p < 0.001, 95% CI [-0.59, -0.31]. In the efficacy analysis, the pre-post effect size of the NOT Fb group was g = -0.42, p < 0.001, 95% CI [-0.58, -0.26]. The effect size for the NOT P/T Fb was g = -0.68, p < 0.001, 95% CI [-1.04, -0.32]. The effect size for the CST Fb group was g = -0.82, p < 0.001, 95% CI [-0.98, -0.314]. The summary of effect sizes from both the ITT and efficacy analyses are presented in Table 7.

Effect sizes for the between-group difference in pre-post change scores on OQ total scale scores in the ITT analyses were as follows. The combined effect size for NOT Fb versus NOT TAU was g = -0.28, p < 0.001, 95% CI [-0.44, -0.11]. The combined effect size for NOT P/T Fb

versus NOT Fb was g = -0.17, p =0.096, 95% CI [-.37, 0.03]. The combined effect size for CST Fb versus NOT Fb was g = -0.21, p = 0.013, 95% CI [-0.38, -0.04]. As in the case with between-group comparisons of mean post-treatment OQ total scores, a mega-analytic approach was utilized to calculate the effect sizes of the CST Fb and the NOT P/T Fb groups in relation to NOT TAU. The effect size for the pooled CST Fb versus pooled NOT TAU was g = -0.43, p < 0.001, 95% CI [-0.58, 0.32]. The effect size for the pooled NOT P/T Fb versus NOT TAU was g = -0.44, p < 0.001, 95% CI [-0.62, -0.27].

In the efficacy analyses, the combined effect size for NOT Fb versus NOT TAU was g = -0.60, p < 0.001, 95% CI [-0.81, -0.40]. When the NOT P/T Fb group was compared to NOT Fb in the efficacy analysis, the combined effect size was g = -0.15, p =0.173, 95% CI [-0.37, 0.07]. The comparison between the CST and NOT Fb groups yielded a combined effect size of g = -0.29, p < 0.062, 95% CI [-0.60, 0.02]. When study 3 was removed from the analysis, the combined effect size was g = -0.40, p = 0.020, 95% CI [-0.74, -0.06]. Utilizing a mega-analytic approach, the effect size between the pooled NOT P/T Fb versus pooled NOT TAU was g = -0.56, p < 0.001, 95% CI [-0.74, -0.37]. The combined effect size between the pooled CST Fb group versus the pooled NOT TAU group was g = -0.78, p < 0.001, 95% CI [-0.96, -0.60].

Effects of feedback interventions on session attendance. The number of therapy sessions utilized by patients was thought of as an effect of feedback interventions in previous studies. Because the number of sessions attended by patients was part of the efficacy criteria, between-group comparisons of the mean number of session attendance were appropriate only for the ITT analyses. The summary of effect sizes and the results of tests of publication bias are presented in Table 6.

Feedback (Fb) effect. A one-way ANOVA was conducted on the NOT Fb groups to test for heterogeneity of effect sizes, which resulted in significant study effect, $F(5, 421) = 2.78$, $p = 0.017$. Although speculations could be made about the presence of possible moderators (e.g., treatment settings in which original studies took place), the data was pooled across studies, given that the difference between the highest mean attendance ($M = 10.8$) and lowest mean attendance ($M = 8.44$), as tested by independent samples *t*-test, was not significant, $t(116) = 1.93$, $p = 0.056$. The one-way ANOVA on the NOT TAU group yielded significant study effect, $F(3, 314) = 6.55$, $p < 0.001$, with a significant difference between the highest mean attendance ($M = 11.22$; study 4) and the lowest mean attendance ($M = 6.03$; study 1), $t(61) = -2.563$, $p = 0.013$. In this case, given such a large discrepancy in mean session attendance, the presence of moderator(s) might have contributed to the heterogeneity. Study 1, in particular, was the only study resulting in a very large effect size, while other studies yielded small effect sizes. The causes for such wide dispersion are not known at this time. Thus, although the data was pooled in favor of ecological validity in this study, future investigation of moderators appears warranted. The combined effect of the differences in mean session attendance between the NOT Fb and NOT TAU groups was $g = 0.27$, $p = 0.217$, 95% CI [-0.16, 0.69] with individual effect sizes ranging from –0.10, $p = 0.459$, 95% CI [-0.37, 0.17] to $g = 1.09$, $p < 0.001$, 95% CI [0.58, 1.60] (Table 19 in Appendix D). The results did not show statistically significant difference in the mean number of sessions utilized between the Fb and TAU groups.

Patient/therapist feedback (P/T Fb) effect. The results of one-way ANOVA on the NOT P/T Fb groups did not support the presence of heterogeneity among mean session attendance across studies, $F(2, 219) = 2.67$, $p = 0.071$. The combined effect size of differences in mean session attendance between the NOT P/T Fb and NOT Fb groups was $g = 0.12$, $p = 0.311$, 95%

CI [0.11, 0.35] with individual effect sizes ranging from g = –0.22, p = 0.356, 95% CI [-0.69, 0.25] to g = 0.23, p = 0.145, 95% CI [-0.08, 0.54] (Table 20 in Appendix D). The results did not support the presence of increase in mean number of sessions in P/T Fb groups. The effect size of the pooled NOT P/T Fb group in relation to the pooled NOT TAU group was g = 0.40, p < 0.0001, 95% CI [0.23, 0.58], indicating the attendance of 2.6 more sessions by those in the P/T Fb group (of which 2.5 sessions occurred after the signal alarm event).

Clinical support tools feedback (CST Fb) effect. One-way ANOVA on the CST Fb groups resulted in significant study effect, $F(2, 412) = 4.50$, p = 0.012. The combined effect of the difference in mean session attendance between the CST Fb and NOT Fb groups was g = 0.41, p = 0.024, 95% CI [0.05, 0.76] with individual effect sizes ranging from g = 0.22, p = 0.106, 95% CI [-0.05, 0.48] to g = 0.816, p < 0.001, 95% CI [0.48, 1.16] (Table 21 in Appendix D). Study 3 resulted in significantly larger number of session attendance (mean difference of 4.6 sessions). Given the possible bias reflected in the assignment process in study 3, another weighted mean effect size was calculated after removing the data from study 3. The result was significant at the .05 level, g = 0.22, p = 0.020, 95% CI [0.035, 0.410], suggesting significantly more average session attendance (1.8 more sessions) by NOT patients in the CST Fb groups than those in the Fb groups on average in routine care. The effect size of the pooled CST group in relation to the pooled TAU group was g = 0.48, p < 0.001, 95% CI [0.33, 0.63], indicating the attendance of 3.4 more sessions[10] by those in the CST group (of which 2.7 sessions occurred after the signal alarm event).

Feedback effects on on-track patients. The effects of P/T Fb and Fb interventions on on-track (OT) patients were tested, using mega-analytic approaches. Prior to comparing

[10] The results were equivalent when the data of patients in study 3 were removed from the analysis, g = 0.45, p < 0.001, 95% CI [0.29, 0.60], or 3.2 more sessions attended by CST Fb group.

feedback intervention groups against TAU, heterogeneity of effects was tested in the same manner as the NOT samples. Statistically significant heterogeneity was detected at the .05 level in both (1) ANCOVAs of mean post-test OQ scores by study with pre-test OQ score as a covariate and (2) ANOVAs of mean pre-test scores by study. The primary reason for this heterogeneity was the significantly higher mean post-test and pre-test scores of patients in study 4 (Hawkins, et al., 2004). When the data of patients from study 4 was removed from the analyses, study effects no longer reached statistical significance. When the data was pooled by treatment conditions across studies and tested for equivalence in mean pre-test scores by independent samples *t*-tests, no significant differences were found. Considering the fact that patients from study 4 were found in all three of the OT treatment conditions (i.e., OT Fb, OT P/T Fb, and OT TAU groups), and based on the equivalent mean pre-test scores by treatment conditions, the data was pooled by OT treatment conditions in favor of ecological validity.

Feedback effects (Fb). Contrary to the findings in a previous meta-analysis (Lambert, et al. 2003), when the mean numbers of sessions attended by patients in OT Fb and OT TAU were compared in ITT analysis, no statistically significant difference was found, $g = 0.01$, $p = 0.792$, 95% CI [-0.06, 0.07]. Although the overall decrease in session attendance was not observed, patients in the OT Fb group, on average, experienced greater therapeutic gains. In terms of mean post-test OQ score difference, ITT analysis was significant at .05 level, $g = -0.12$, $p < 0.001$, 95% CI [-0.19, -0.06], equivalent of approximately 2.8 OQ point reduction, while efficacy analysis was also significant at .05 level, $g = -0.30$, $p < 0.001$, 95% CI [-0.37, -0.23], equivalent of 6.5 OQ point reduction.

When the odds of patient deterioration/reliable worsening at termination of the OT Fb group were compared against OT TAU, the results of ITT mega-analyses indicated that the effect

size was significant at the .05 level, *OR* = 0.63, *p* = 0.030, 95% CI [0.41, 0.95]. Efficacy analysis was not significant at the .05 level, *OR* = 0.81, *p* = 0.341, 95% CI [0.52, 1.25]. When the odds of the occurrence of patient reliable/clinically significant improvement at termination were compared, both ITT and efficacy analyses were significant at the .05 level with respective odds ratios of *OR* = 1.20, *p* = 0.010, 95% CI [1.04, 1.38] and *OR* = 2.09, *p* < 0.001, 95% CI [1.80, 2.42]. These results suggest that, while the amount of utilization of sessions essentially remains the same, OT patients in Fb condition on average experience superior treatment outcome and may have decreased odds of experiencing deterioration than those in TAU.

Patient/therapist feedback effects (P/T Fb). When the mean number of session attendance was compared between the OT P/T Fb and OT TAU groups in ITT analysis, the results were significant at .05 level, *g* = 0.10, *p* < 0.02, 95% CI [0.01, 0.18], equivalent of approximately 0.4 more sessions attendance by the OT P/T Fb. In terms of mean post-test OQ score differences, both ITT and efficacy analyses respectively yielded significant results at the .05 level, *g* = -0.18, *p* < 0.001, 95% CI [-0.26, -0.96], equivalent of approximately 4.1 OQ point reduction on average, and *g* = -0.32, *p* < 0.001, 95% CI [-0.40, -0.23], equivalent of approximately 7.1 OQ score point reduction on average.

When the odds of the occurrence of patient reliable worsening/deterioration at termination were compared, both ITT and efficacy analyses were not significant at the .05 level with respective odds ratios of *OR* = 0.71, *p* = 0.215, 95% CI [0.42, 1.22] and *OR* = 0.86, *p* = 0.585, 95% CI [0.55, 1.47]. When the odds of the occurrence of reliable/clinically significant improvement were compared, ITT and efficacy analyses respectively yielded significant results at the .05 level, *OR* = 1.65, *p* < 0.001, 95% CI [1.39, 1.96] and *OR* = 2.36, *p* < 0.001, 95% CI [1.97, 2.82]. These results suggest that, in comparison to TAU, patients who receive P/T Fb

intervention on average experience superior treatment outcome in terms of distress reduction and improved odds of achieving reliably positive change, while the odds of reliable worsening/deterioration remain the same as those for patients in TAU.

Discussion on Study 1

This meta- and mega-analytic study evaluated the effects of three types of patient progress feedback interventions used in the OQ-based quality assurance system: progress feedback to therapists, progress feedback to both patients and therapists, and Clinical Support Tools in addition to progress feedback. These interventions were aimed at monitoring individual patient progress in treatment, identifying patients at risk of treatment failure, and intervening before termination occurs. The effects of these interventions were evaluated with patients whose progress in treatment was identified as not-on-track (NOT), i.e., at risk of leaving treatment worse off than when entering the treatment as well as those identified as on-track (OT). Two sets of analyses were conducted to estimate the effects of feedback interventions that can be expected in routine practice (intent-to-treat analyses; ITT) and among patients who stay in treatment until the effects of feedback interventions could be measured (efficacy analyses). The effects of the feedback interventions were evaluated on the basis of between group differences in mean OQ total scores at termination of treatment, rate and odds of clinically significant change status at termination, and mean number of sessions attended.

Overall, the effects of feedback interventions on patients who were identified as being at risk of treatment failure (NOT) were more substantial than those identified as being on-track. When compared to the NOT TAU in ITT analyses, the combined effects (Hedges's *g*) of mean post-treatment OQ total scores for the NOT Fb, NOT P/T Fb, and the CST Fb groups were -0.28, -0.36, and -0.44, respectively. The overall percentages of reliable worsening/deterioration

(clinically significant improvement) among the NOT TAU, NOT Fb, NOT P/T Fb, and CST Fb groups were 20.1% (22.3%), 13.6% (30.9%), 15.8% (38.7%), and 11.3% (37.6%), respectively. The odds ratio of reliable worsening/deterioration (clinically significant improvement) for the NOT Fb, NOT P/T Fb, and CST Fb groups in relation to NOT TAU were 0.62 (1.70), 0.74 (2.20), and 0.51 (2.01), respectively. These results indicate that all forms of feedback interventions were effective in enhancing outcome while reducing treatment failures among NOT patients, with an exception of the P/T Fb intervention in its effects in preventing treatment failure. These results also show that, when the treatment impact is evaluated on the level of routine care (ITT analysis), the three types of feedback interventions are similar in their effects on treatment outcomes.

The effects of feedback interventions on those who satisfied the least necessary conditions to likely have been the actual recipients of the feedback interventions were also estimated (efficacy analyses). Such criteria comprised of minimum numbers of session attendance (i.e., at least four to six sessions, depending on feedback conditions) and minimum numbers of completion of the OQ (i.e., at least three to four administrations, depending on conditions). The effect sizes (Hedges's *g*) for the mean post-treatment OQ total score differences between the NOTFb, NOT P/T Fb, and CST Fb groups in comparison to NOT TAU were -0.53, -0.55, -0.70, respectively. Furthermore, the percentages of patients experiencing reliable worsening/deterioration (clinically significant improvement) for the NOT TAU, NOT Fb, NOT P/T Fb, and CST Fb groups were 20.1% (22.3%), 9.1% (37.6%), 14.7% (45.2%), and 5.5% (52.5%), respectively. The combined odds ratio of reliable worsening/deterioration (clinically significant improvement) for the NOT Fb, NOT P/T Fb, and CST Fb groups were 0.44 (2.55), 0.68 (2.97), and 0.23 (3.85), respectively. These results indicate greater treatment effects

in all of the outcome criteria evaluated in this study, except for the NOT P/T Fb condition in its effect to reduce reliable worsening/deterioration.

Contrary to the previous meta-analysis (Lambert, et al., 2003), this study highlighted effects of feedback interventions on OT patients. Although not to the magnitude experienced by the NOT counterparts, patients in OT P/T Fb and OT Fb appear to experience more distress reduction and increased odds of experiencing reliable/clinically significant improvements than those in OT TAU.

It is interesting to note the pattern of outcomes seen in the P/T Fb patients. Specifically, this intervention yielded increased treatment enhancing effects while yielding similar rate of reliable worsening/deterioration when compared with that of the TAU group. These results suggest the possibility of a mechanism that interacts with provision of direct progress feedback to patients in such a way that enhances outcome for some, while inhibiting outcome enhancement for others.

Another aim of this meta-analysis was to investigate the incremental benefit of the NOT P/T Fb and CST Fb interventions in comparison to the NOT Fb intervention. Because previous reports (Harmon, et al., 2007; Slade, et al., 2008) provided only the results of efficacy analyses for the CST Fb intervention, I compared this intervention against the NOT Fb intervention under equivalent inclusion criteria. Although the comparative magnitude of the effects of the CST Fb over the Fb were smaller than previously reported (primarily due to comparing the efficacy CST Fb samples against the ITT NOT Fb samples in previous studies), results of ITT analyses produced statistically significant effects in terms of superior distress reduction at post-treatment (g = -0.16, $p < 0.05$) and increased odds of clinically significant improvement (OR = 1.53, $p <$ 0.05). Although statistical significance was observed in comparisons between the CST Fb and

NOT Fb groups, there were overlaps in 95% confidence intervals between the CST Fb versus NOT TAU comparisons and the NOT Fb versus NOT TAU comparisons. Efficacy analyses, however, did not yield statistically significantly greater additive treatment effects for the CST Fb group over the NOT Fb group. These statistically non-significant results, however, should not be automatically assumed to be indications of *no* additive effect, as reflected in greater effect sizes yielded in the efficacy analyses. Statistically non-significant results may be due to the lack of statistical power because of the reduction of sample sizes by applying exclusion criteria. Future trials featuring the CST Fb and P/T Fb interventions may further our understanding of the magnitude of these interventions.

Comparison of ITT analyses and efficacy analyses opens questions about possible mechanisms of change. Because the primary element of the exclusion criteria used in efficacy analyses was the number of sessions attended by patients, a question arises as to how much of the improvements in the results of efficacy analyses were function of the dose-response effect (Hansen, Lambert, & Forman, 2002). Improvements in treatment outcome among the efficacy samples also suggest a higher proportion of poorer outcomes among patients who left treatment before the feedback interventions could have taken the effect. If so, it appears important that proactive effort be given to retain at risk patients in treatment, even more so for those experiencing worsening at early stages of therapy.

As a supplemental analysis, to test the possibility of disproportional occurrences of reliable worsening/deterioration and clinically significant improvement based on the length of treatment, the percentages and odds of such outcomes on the pooled dataset of all NOT cases (N = 1382) were calculated. Of those NOT patients who left treatment after five sessions or less (early terminators; n = 381), 22.8% deteriorated, while 12.1% clinically significantly improved.

In contrast, of those NOT patients who stayed in treatment six sessions or more (late terminators; *n* = 1001), 11.7% later reliably worsened/deteriorated, while 39.9% clinically significantly improved. When early terminators and late terminators were compared, the odds ratio of reliable worsening/deterioration for the early terminator was 2.24, while the odds ratio of clinically significant improvement was 0.21. These findings underline the need to retain NOT patients in treatment longer. Future research effort to uncover the therapeutic and counter-therapeutic processes of engaging NOT patients in treatment is recommended. Future research concerning the process of deterioration also appears to be an important area to be explored further.

Limitations of study 1. While research synthesis, such as this quantitative review, provides various statistical advantages in data analyses, this study has limitations, many of which were inherent in the original studies. Reliability of treatment implementation may have been an issue in individual studies as the use of feedback interventions by therapists were not closely controlled or monitored. While statistical power increased owning to data synthesis/pooling, the magnitude of true effects may have been underestimated. Because random assignment to conditions was not incorporated in two of the studies (Lambert, Whipple, Vermeersch et al., 2002; Whipple, et al., 2003), selection bias may have occurred, resulting in heterogeneous samples of patients. Similarly, an argument can be made against causal statements based on the data of studies not directly compared in the same randomized trials, such as in the case of comparing the CST Fb group and the pooled TAU group in the Harmon and Slade studies. However, pooled TAU data from multiple studies may provide the most reliable benchmark for comparing alternative treatment strategies.

Although this issue was not detected in original studies, application of uniform inclusion/exclusion criteria in this study revealed some heterogeneity of effects across studies.

However, examination of evidence based on different inclusion criteria shed some light on the consequences of applying such criteria. Through the use of intent-to-treat and efficacy analyses, this study provided better understanding about the effects that can be expected when the feedback interventions are implemented as a policy, as well as the effects expected from controlled trials. Nonetheless, further replications across different patient populations by different research groups are needed before the boundary conditions of effectiveness will be known.

Another criticism may be made regarding the possibility of mono-method bias because this line of research used only the OQ-45 as the outcome measure as well as the method for identifying NOT cases. An argument can be made of using multi-method, multi-perspective outcome assessment to capture more breadth of information related to patient treatment progress and outcome. Such methods may be valuable in enhancing more comprehensive understanding of the impact of feedback interventions than the methodology utilized in the line of research reviewed. However, routine assessment and monitoring of outcome requires an instrument that is time- and cost-efficient. In routine care where treatment termination is determined largely by patients and treatment length is unknown at the outset, the use of multiple outcome measures is not feasible. Given the established reliability and validity of the OQ-45 as a sensitive measure of treatment outcome (Vermeersch, et al., 2004) and that its' classification of patient change is concordant with other frequently used measures, the use of the OQ-45 as the sole assessment tool seems well suited for the purpose of quality assurance in routine clinical practice.

Another limitation of this line of research is the exclusive use of OQ total score in outcome monitoring and feedback provision. Although extensive examination of OQ subscales was performed in this study, the original feedback studies did not utilize subscale information in

progress feedback. Thus, the treatment effects reflected at the subscale level were still effects based on the OQ total score system. As already discussed in the meta-and mega-analytic investigations of the OQ subscale, subscale based information could provide unique clinical information that cannot be captured by the total scale score alone. Development and implementation of OQ subscale score based progress feedback, in addition to the well established total scale score based system may further enhance the clinical utility of this quality assurance system. One practical advantage of subscale-based feedback system is that it requires no more burden on patients other than filling out the OQ as they normally do.

Although I do not considered the following as a limitation to this line of research, it is important to point out that the feedback procedures in the OQ quality assurance system appear to be more appropriate for cases that are predicted to deteriorate, not all patients. To better understand the effects of feedback interventions in a broad context of routine clinical practice, the overall effects of feedback interventions on all patients included in original studies (both on-track and not-on-track patients) by pooling the entire datasets of the six studies (N = 6151; ITT analyses). Of those who received any form of feedback interventions, 4.7 % of patients experienced reliable worsening or deterioration, while 37.4% of patients experienced clinically significant improvement. Of those patients in routine care (i.e., no feedback; treatment-as-usual), 6.1% reliably worsened/deteriorated while 30.2% achieved clinically significant improvement. Accordingly, overall odds of deterioration among the pooled feedback interventions group in relation to patients receiving treatment-as-usual were statistically significant (OR = 0.76, p = 0.024). The overall odds of clinically significant improvement among those in the pooled feedback group was also statistically significant (OR = 1.38, p < 0.001). The overall effects in terms of post-treatment mean OQ total scores showed significantly

less disturbance (g = -0.12, SE = 0.03, $p < 0.001$). This effect size translates to 2.9 OQ total points reduction on average. The overall reduction of deteriorated cases, increase in clinically significant improvement, and decrease in distress level at termination occurred within a context of utilizing an average of 0.9 more session of care. Although the value judgment regarding the average increase of 0.9 more sessions of care may vary, the benefit of the feedback interventions, especially for patients who are at risk of treatment failure, appears worthy of a serious consideration for routine implementation of outcome monitoring and provision of feedback.

Study 2: Meta-analysis of Outcome Questionnaire-45 Subscale Scores

Selection Criteria and Participants

Retrieval of OQ subscale scores. The effects of feedback interventions on OQ subscale scores were investigated, employing the same statistical methodologies used in the meta-analysis and mega-analysis of the feedback effects on the OQ total scale scores. Prior to conducting statistical analyses, however, significant effort was expended on retrieving the OQ subscale scores for each of the datasets of the original feedback studies because OQ subscale scores were not included in the original datasets. Brigham Young University Counseling and Career Center, where five of the six studies in this review were conducted, maintained a separate dataset of OQ scores for research purposes. The person ID number, date of OQ administration, and OQ total scores were used to retrieve the OQ subscale scores for each data point for each participant. The original dataset of study 4 (Hawkins et al., 2004) no longer contained the participants' original ID numbers, thus, the matching procedure became difficult. To make the procedure further complicated, study 4 had no backup dataset with original OQ subscale scores, thus, the hard copy of the administered OQ protocols were re-scanned, cleaned, and scored prior to commencing the matching procedure. Due to various human errors inherent in filling out paper forms, complete

replication of the original dataset based on the re-scanned data was not achieved. In study 4, each individual data point was matched on the basis of the combinations of demographic variables that were common to both the original dataset and the re-scanned dataset, such as therapist ID, date of OQ administration, OQ total score, patient gender, marital status, and employment status, in so far as they were available.

Inclusion and exclusion criteria. 6,151 patients included in the meta-analysis of OQ total scale scores utilized a total of 33,558 sessions. Of these sessions recorded in the original feedback study datasets, patients as a whole completed the OQ-45 in 91.4% of the sessions, which resulted in 30,660 data points with OQ administrations. Of these, 29,437 observations were matched with subscale scores, using series of syntaxes that outlined variables to link between each observation in the datasets from the original studies to corresponding subscale scores. Some subscale scores could not be matched by the use of syntaxes alone because of some presumed human errors at the time of data entry, such as dates of the OQ administration entered incorrectly in either the original dataset or the subscale score database. Judging the relations to other variables, however, an additional 709 observations were matched manually. In sum, 30,146 sessions out of 30,660 sessions (98.3%) succeeded in retrieval of the OQ subscale scores (N = 6044).

The accuracy of the data match was assessed in two ways. First, for each case, the total scale scores from the original studies and those derived from the matched datasets were compared at pre-treatment, the time of signal warning (for not-on-track cases), and post-treatment. The following number of cases had discrepancies by greater or smaller than four OQ points between the total scale scores from the original studies and those derived from the merged datasets at pre-treatment (n_1), the time of warning (n_2), post-treatment (n_3), and both pre-

treatment and post-treatment (n_4), respectively: $n_1 = 261$, $n_2 = 45$, $n_3 = 206$, and $n_4 = 25$. In some cases, discrepancies in scores seemed so large that the two scores being matched seemed to come from different data points. However, given that the reasons for such discrepancies were unknown, all merged data was included in this study. This decision to include all of the merged data has some possible consequences. For instance, discrepancies in pre- or post-treatment total scores could alter such relationships in the data as the correlations between the pre-post change scores in the total scale and subscales. Clinical significance classifications, which also rely on the pre-post change scores as well as the pre- and post-treatment scores, may not match depending on the degree of discrepancies. In the second approach to assessing the accuracy of the data match, the mean, standard deviation and *n* of OQ total scale scores from the original study datasets and merged datasets were calculated and compared at pre-treatment, time of signal warning, and post-treatment (Table2.1). As can be seen on Table 9, the merged data appears to have accurately reproduced the original data structures.

Assessment of relations between OQ total scale and subscales. As discussed earlier, OQ subscale scores have not been studied in previous feedback studies. Thus, to interpret the results of analyses based on OQ subscale scores, gaining rudimentary understanding about the relations between the OQ total scale scores and subscale scores in the feedback studies was deemed necessary. Several approaches were utilized to explore such relations. First, the correlations between the total scale score and each of the subscale scores at three time points were calculated: pre-treatment, time of signal warning, and post-treatment. These correlations were thought of as initial estimates of similarities between the effect sizes obtained from group differences in mean post-treatment OQ total scale scores and effect sizes obtained from corresponding subscale scores. Second, the correlations between the pre-post change scores

between the total scale and subscales were calculated to obtain the estimates of associations between the total scale and subscale pre-post change effect sizes. Third, to investigate the relations between the clinical significance status of the outcome at termination based on the total scale scores and those based on subscale scores, flow charts were created to show the breakdown of the n and percentage of matched/non-matched cases.

Correlation between OQ total scale scores and subscale scores. The Pearson's correlation coefficients between the total scale and subscale scores at pre-treatment, signal warning, and post-treatment as well as pre-post change scores are reported in Table 10. High correlations were found between the OQ total scale scores and the SD scale scores. The IR and SR subscales presented lower, but still moderate correlations. Correlation coefficients of the IR and SR subscales were very similar to each other. These results suggested that the meta- and mega-analytic results in this study on the SD subscale would be similar to those found in the total scale analyses.

Match in clinical significance status between OQ total scale scores and subscale scores. To investigate the matches between clinical significance status between the outcome classification system based on the total scale and those based on the subscales, patient outcome status at termination were broken down to the following categories. To facilitate ease in understanding various classifications, a flowchart of patient classification is provided in Figure 1.

It should be noted, however, that these categorizations are not exhaustive as to including all possible combinations between the classifications based on the total scale scores and those based on the subscale scores.

- Deterioration cases—(a) patient was classified as a deterioration case based only on the total scale, but not on any of the subscales; (b) patient was classified as a deterioration case based on both the total scale and at least one subscale;
- No change cases—(c) patient was classified as achieving no significant change on both the total scale and all of the subscales, (d) patient was classified as achieving no significant change on the total scale, but was identified as a deterioration case on at least one subscale, (e) patient was classified as achieving no significant change on the total scale, but was identified as a reliable improvement/clinically significant improvement case on at least one subscale, (f) patient was classified as a no change case on the total scale, but had mixture of both deterioration and improvement on subscales.
- Reliable improvement/clinically significant improvement—(g) patient was classified as a reliable improvement/clinically significant improvement case on the total scale only, and (h) patient was classified as a reliable improvement/clinically significant improvement case on both the total scale and at least one subscale.

In a small minority of cases, outcome classifications in the total scale and subscales were contradictory (e.g., clinically significant improvement on the total scale, but one of the subscales indicated deterioration). In approximately half of these cases, the cause of such discrepancy seemed to have resulted from discrepancies in the data match as described earlier. In other cases, the discrepancy in outcome classification seemed to have resulted from large changes on the SD subscale "masking" the changes of other subscales that went in the opposite direction. As presented in Figure 1, cases where both the total scale score-based clinical significance classification and those based on the subscales matched, *n* and percentage of the given clinical

significance (i.e., deterioration, no change, or reliable improvement/clinically significant improvement) on each subscale and combinations of subscales are reported. The results show about 79% of deterioration cases based on the total scale classification were also identified as deterioration cases in at least one of the subscales. Nearly 79% of those subscale-based deterioration cases involved either deterioration in the SD subscale alone (43%) or combinations of the SD scale and other subscales. Interestingly, deterioration on the IR and SR subscales were identified mostly when they were the sole subscale classified as deterioration (10% each) or when combined with the SD scale. Only in rare instances (1.6%) did deterioration occur on the combination of the IR and SR scales, but not the SD scale, when deterioration was also observed on the total scale.

The matching results further showed that when patients were classified as reliable improvement/clinically significant improvement cases, in approximately 90% of the time patients were also classified as achieving the same outcome on at least one subscale or combinations of subscales. In nearly 95% of those matched cases involved clinically significant improvement in the SD subscale, either as the SD scale alone (48%) or the SD subscale in combination with other subscales (47%). The occurrences of clinically significant improvement on the IR and SR subscales seemed underrepresented, especially when they were the sole subscale being identified as meeting the criteria for achieving reliable improvement/clinically significant improvement (2 to 3%).

These results initially seemed to suggest close overlaps between the total scale-based outcome classification and the SD subscale-based classification. However, a closer look at the data of “no change” cases on the total scale revealed that the IR and SR subscale-based classifications provided unique outcome information that was not detected by the total scale-

based system. For instance, 166 patients in the no-change category based on the total scale outcome classification (equivalent of more than a half the number of those who were identified as deterioration cases based on the total scale system, *n* = 309) experienced deterioration on at least one subscale. Combining all deterioration cases based on the IR subscale, 37% (*n* = 59) were classified as no-change cases on the total scale outcome classification. Similarly, of all the SR subscale-based deterioration cases, 33% (*n* = 46) of them were identified as no-change cases on the total scale classification. 25% of the deterioration cases on the SD subscale were classified as no-changers on the total scale.

These results suggest that a sizable portion of deterioration at the subscale level was not detected by the total scale score classification. In contrast, relatively smaller percentage of improvement cases based on the subscale classification systems were identified as no change cases on the total scale system. Of all the improvement cases on the subscales, 12% from the SD subscale, 6% from the IR subscale, and 7% from the SR subscale (including combinations of clinically significant improvement status attained on multiple subscales) were classified as no change cases on the total scale. These results suggest that when clinically significant improvement occurred at the subscale level, such outcome was likely reflected in the total scale score based outcome classification. These results appeared to support the possible unique contributions the OQ subscale analyses might add to our understanding of the effects of feedback interventions on patient outcome. It is important to note that in the original feedback studies, the feedback provided to clinicians regarding their patients' progress was based on the OQ total scale scores alone. It is thus possible that the provision of OQ subscale based feedback might have resulted in greater change on the outcome at the subscale level. *The analyses reported in this review, however, examined the effects of the OQ total scale scores-based feedback*

interventions as measured at the OQ subscale level, not the effects of providing feedback based on the OQ subscale scores.

Dependent Measures and Computation of Effect Sizes

To be consistent with the methodologies utilized in the OQ total scale analyses, the following dependent measures were used to assess the effects of feedback treatments: Mean post-treatment OQ subscale scores, mean pre-treatment to post-treatment change scores, the odds of patients achieving clinically significant improvement at post-treatment based on the subscale-level outcome classification, and the odds of the occurrence of clinically significant worsening (or deterioration) at post-treatment based on the subscale-level outcome classification. Effect sizes were calculated based on the methodologies utilized for the OQ total scale analyses, except as noted in specific sections below.

Analysis of OQ Subscale Effect Sizes

Post-treatment between-group difference. In the sections below, the meta-analytic and mega-analytic results of post-treatment between-group differences in each of the OQ subscales are discussed in turn. The results of each subscale analysis, including the results of tests of publication bias, are presented in Tables 11, 12, and 13. To facilitate ease in comparing effect sizes between groups and across different subscales, effect sizes and corresponding 95% CIs are summarized in Table 14.

Symptom distress (SD) scale.

Feedback effect. The results of one-way ANCOVA, testing for heterogeneity of effects across studies, with study as the factor, post-treatment SD subscale score as the dependent variable, and pre-treatment SD subscale score as the covariate, indicated no significant study effect among the Fb group in ITT analysis, $F(3, 261) = 0.88$, $p = 0.883$, or efficacy analysis $F(3,$

130) = 1.43, p = 0.237. Similarly, no statistically significant study effect was found among the NOT TAU, $F(3, 307) = 1.88$, p =0.133. These results do not support the presence of heterogeneity of effect sizes. The results of independent samples t test, comparing the pre-treatment symptom distress level as measured by the SD subscale between the pooled NOT Fb and the NOT TAU groups showed no between treatment group difference in either ITT analysis, $t(576) = -0.55$, p = 0.586, or efficacy analysis, $t(445) = -0.33$, p = 0.740.

The combined effect size for the mean difference in post-treatment SD subscale scores between the NOT Fb group and the NOT TAU group was significant in ITT analysis, g = -0.26, p = 0.002, 95% CI [-0.42, -0.09], equivalent of 3.7 OQ points difference in mean post-treatment SD scores, with individual effect sizes ranging from g = -0.38, p = 0.137, 95% CI [-0.88, 0.12] to g = -0.04, p = 0.878, 95% CI [-0.53, 0.46]. The combined effect size for the efficacy analysis was also significant, g = -0.52, p < 0.001, 95% CI [-0.73, -0.32], equivalent of 7.4 OQ points difference, with individual effect sizes ranging from g = -0.69, p < 0.001, 95% CI [-1.02, -0.36] to g = -0.27, p = 0.337, 95% CI [-0.82, 0.28].

Patient/therapist feedback (P/T Fb) effect. As with the meta-analyses of the OQ total scale scores, incremental effects of the NOT P/T Fb and CST Fb interventions were tested against the NOT Fb group. The results of ANCOVAs, testing the heterogeneity of effects by study with pre-treatment SD subscale scores as a covariate did not reach statistical significance at the .05 level in ITT analysis, $F(2, 217) = 1.46$, p = 0.234, or efficacy analysis, $F(2, 172) = 1.56$, p = 0.212. Mean pre-treatment SD subscale scores were near identical between the NOT P/T Fb group and NOT Fb group in both ITT analysis, $t(405) = -0.11$, p = 0.912 and efficacy analyses, $t(321) = -0.17$, p = 0.865. The combined effect size for the comparison between NOT P/T Fb and NOT Fb was not significant in either ITT analysis, g = -0.11, p = 0.272, 95% CI [-0.30,

0.09], or efficacy analyses g = -0.09, p = 0.403, 95% CI [-0.31, 0.13]. These results suggest that provision of progress feedback to both patients and therapists adds little benefit to provision of feedback to therapists alone in reducing symptom distress.

The NOT P/T Fb group was also compared to the NOT TAU group. The results for independent samples t-test of pre-treatment SD subscale scores between pooled NOT P/T Fb and pooled NOT TAU groups was not significant in either ITT analysis, $t(531) = 0.483$, p = 0.629, or efficacy analysis, $t(486) = 0.21$, p = 0.838. These results indicate equivalence of pre-treatment distress level between the two groups. The effect size in ITT analysis was g = -0.24, p = 0.008, 95% CI [-0.41, -0.06], an equivalent of 3.3 OQ point difference on average, favoring the NOT P/T Fb group over the NOT TAU group. The effect size in efficacy analysis was g = -0.39, $p <$ 0.001, 95% CI [-0.57, -0.20], equivalent of 5.4 OQ point difference on average, favoring NOT P/T Fb group. These results show that the effect sizes of post-treatment comparisons between the NOT P/T Fb group and the NOT group are smaller than those between the NOT Fb group and the NOT TAU. Yet, the effect sizes between the NOT P/T Fb and NOT Fb were in favor of the NOT P/T Fb albeit the results did not reach statistical significance.

These seemingly inconsistent findings were the results of having two different NOT Fb groups in the above comparisons. In the case of comparisons between the NOT Fb and NOT TAU, studies 1, 2, 3, and 4 (i.e., Lambert, et al., 2001; Lambert, et al., 2002; Whipple, et al., 2003; and Hawkins, et al., 2005) were used. The comparisons between the NOT P/T Fb and NOT Fb groups were based on studies 4, 5, and 6 (i.e., Hawkins, et al., 2005, Harmon, et al., 2007, and Slade, et al., 2008). The NOT Fb groups from the latter studies had a higher mean post-treatment scores (i.e., higher distress) than those from the earlier studies.

Clinical Support Tools feedback (CST Fb) effects. The results of one-way ANCOVA to test for the presence of heterogeneity of effects resulted in a statistically significant result among the CST group in ITT analysis, $F(2, 408) = 4.97$, $p = 0.007$, but not in efficacy analysis, $F(2, 210) = 0.264$, $p = 0.768$. The significant results seen in the ITT analysis appears mainly due to study 3 (Whipple, et al., 2003), where the mean pre-treatment SD subscale score was higher and post-treatment SD subscale score was lower than other studies. Controlling for the pre-treatment score appears to have contributed to a significant discrepancy on the mean post-treatment SD subscale scores. The combined effect size for the group difference between the CST Fb and NOT Fb groups at post-treatment was not significant for the ITT analysis, $g = -0.15$, $p = 0.066$, 95% CI [-0.32, 0.01], with individual effect sizes ranging from $g = -0.07$, $p = 0.623$, 95% CI [-0.33, 0.20] to $g = -0.24$, $p = 0.082$, 95% CI [-0.50, 0.03]. The combined effect size was also not significant in efficacy analysis $g = -0.18$, $p = 0.086$, 95% CI [-0.38, 0.03], with individual effect sizes ranging from $g = -0.29$, $p = 0.085$, 95% CI [-0.62, 0.04] to $g = 0.08$, $p = 0.736$, 95% CI [-0.37, 0.52].

Although statistical significance was not achieved, both the ITT and efficacy effect sizes resembled those found in the OQ total scale scores analyses, including the unique behaviors of the CST Fb group in study3. The lack of statistical power may have been an issue. The CST Fb group was also mega-analytically compared with NOT TAU, comparing the pooled datasets of the two groups. The results of independent samples *t*-test of difference in mean pre-treatment SD subscale scores between the pooled CST Fb and the pooled NOT TAU did not reach statistical significance in ITT analysis, $t(722) = -0.416$, $p = 0.678$ or efficacy analysis, $t(524) = 0.517$, $p = 0.606$, indicating that both groups were comparable in pre-treatment symptom distress level as measured by the SD subscale. The effect size in ITT analysis was $g = -0.33$, $p < 0.001$,

95% CI [-0.47, -0.18], equivalent of 4.5 OQ point difference on average, favoring the CST Fb group. The effect size in efficacy analysis was g = -0.55, $p < 0.001$, 95% CI [-0.73, -0.38], equivalent of 7.5 OQ point difference on average, favoring the CST Fb group.

Interpersonal relations (IR) subscale.

Feedback (Fb) effects. One-way ANCOVAs were performed to test for the heterogeneity of effects, using post-treatment IR subscale score as the dependent variable, study as a factor, and pre-treatment IR subscale score as a covariate. The NOT Fb group did not result in significant study effects at the .05 level for either ITT analysis, $F(3, 261) = 1.12$, $p = 0.342$, or efficacy analysis, $F(3, 130) = 1.72$, $p = 0.165$. The NOT TAU group resulted in significant heterogeneity, $F(3, 307) = 3.18$, $p = 0.024$. The cause of this heterogeneity appears multifaceted, including varying levels of pre-treatment and post-test IR scores across studies as well as directions of pre-treatment to post-treatment change on this subscale. Due to this heterogeneity, the effect sizes were calculated both meta-analytically and mega-analytically, where patients were grouped into pooled NOT Fb or NOT TAU groups. The effect sizes were essentially identical, thus only the meta-analytic results are presented here. There was no significant difference in the mean pre-treatment IR score between NOT Fb and NOT TAU in ITT analysis, $t(576) = -0.92$, $p = 0.357$, or efficacy analysis, $t(445) = -0.72$, $p = 0.473$, indicating that the two groups were comparable in the average level of disturbance in interpersonal relations at pre-treatment.

In ITT analysis, individual effect sizes ranged from g = -0.40, $p = 0.002$, 95% CI [-0.65, -0.14] to g = -0.12, $p = 0.638$, 95% CI [-0.38, 0.62]. The combined effect size for the mean post-treatment difference in mean IR subscale scores between NOT Fb and NOT TAU was g = -0.24, $p = 0.013$, 95% CI [-0.43, -0.05], representing the average of 1.5 IR subscale score

difference on average, favoring the NOT Fb group. In efficacy analysis, individual effect sizes ranged from -0.72, $p < 0.001$, 95% CI [-1.05, -0.38] to $g = -0.03$, $p = 0.926$, 95% CI [-0.58, 0.52]. The combined effect was $g = -0.37$, $p = 0.016$, 95% CI [-0.67, -0.07], indicating an average of 2.4 difference in IR subscale scores at post-treatment, favoring the NOT Fb group. It should be noted that, because the IR and SR subscales have far fewer number of items than the SD subscale, effect sizes of similar magnitude on the IR and SR subscales do not translate to similar mean differences in the raw score unit (i.e., OQ points).

Patient/therapist feedback (P/T Fb) effect. The results of one-way ANCOVA to test for the potential study effects did not support the presence of heterogeneity in either ITT analysis, $F(2, 217) = 0.05$, $p = 0.952$ or efficacy analysis, $F(2, 172) = 0.251$, $p = 0.778$. Mean pre-treatment IR subscale scores were near identical between the NOT P/T Fb group and NOT Fb group in both ITT analysis, $t(405) = -0.07$, $p = 0.944$ and efficacy analyses, $t(321) = -0.10$, $p = 0.924$. In ITT analysis, individual effect sizes in the mean post-treatment IR subscale scores between NOT P/T Fb and NOT Fb ranged from $g = -0.38$, $p = 0.121$, 95% CI [-0.86, 0.10] to $g = 0.06$, $p = 0.671$, 95% CI [-0.23, 0.36]. The combined effect size was $g = -0.06$, $p = 0.569$, 95% CI [-0.28, 0.15]. In efficacy analysis, individual effect sizes ranged from $g = -0.48$, $p = 0.103$, 95% CI [-1.05, 0.10] to $g = -0.01$, $p = 0.958$, 95% CI [-0.35, 0.33] with the combined effect size of $g = -0.12$, $p = 0.274$, 95% CI [-0.34, 0.10]. These results suggest that provision of progress feedback to patients and therapists to NOT patients does not appear to provide unique advantage over provision of progress feedback to therapists alone in improving interpersonally oriented outcome. The NOT P/T Fb group was also compared to NOT TAU, using a mega-analytic approach. The results of independent samples *t*-test did not yield group difference in mean pre-treatment IR scores in either ITT analysis, $t(531) = -0.247$, $p = 0.805$ or efficacy analysis, $t(486)$

= -0.497, p = 0.620, indicating the two groups were comparable at pre-treatment in distress related to interpersonal relations. The effect size in ITT analysis was g = -0.29, p = 0.001, 95% CI [-0.46, -0.12], equivalent of 1.9 OQ points, favoring the NOT P/T Fb. The effect size in efficacy analysis was g = -0.43, p < 0.001, 95% CI [-0.61, -0.24], equivalent of 2.8 OQ points difference on average, favoring the NOT P/T Fb.

Clinical support tools feedback (CST Fb) effects. The results of ANCOVAs did not support the presence of heterogeneity of effects in terms of study effects in either ITT analysis, $F(2, 408) = 2.18$, p = 0.115, or efficacy analysis, $F(2, 210) = 1.04$, p = 0.357. Pre-treatment mean IR subscale scores of CST Fb and NOT Fb were comparable in both ITT sample, $t(654)$ = 0.63, p = 0.512, and efficacy sample, $t(380) = 1.58$, p = 0.116. In ITT analysis, individual effect sizes ranged from g = -0.13, p = 0.459, 95% CI [-0.46, 0.21] to g = -0.05, p = 0.736, 95% CI [-0.31, 0.22], with the combined effect size of g = -0.08, p = 0.357, 95% CI [-0.24, 0.09]. In efficacy analysis, the combined effect size was also statistically non-significant, g = -0.12, p = 0.247, 95% CI [-0.32, 0.08] with individual effect sizes ranging from g = -0.24, p = 0.152, 95% CI [-0.57, 0.09] to g = -0.01, p = 0.952, 95% CI [-0.46, 0.43]. These results suggest that CST Fb intervention has little advantage over NOT Fb in improving interpersonal aspects of outcome in terms of the post-treatment IR scores.

The CST Fb group was also compared to NOT TAU. Pre-treatment mean IR subscale scores did not differ in either ITT analysis, $t(722) = -1.17$, p = 0.242 or efficacy analysis, $t(524)$ = 0.15, p = 0.882, indicating that patients in CST Fb and NOT TAU on average began treatment with equivalent level of interpersonal disturbance. In ITT analysis, the effect size in terms of mean post-treatment IR subscale scores difference was g = -0.42, p < 0.001, 95% CI [-0.51, -0.27], equivalent of 2.7 OQ points difference on average, favoring the CST Fb group.

The effect size in efficacy analysis was g = -0.54, p < 0.001, 95% CI [-0.71, -0.36], equivalent of 3.5 OQ point difference on average, favoring the CST Fb group. These results suggest CST Fb group had a moderate effect in improving interpersonally oriented outcome.

Social role (SR) subscale.

Feedback (Fb) effects. The results of one-way ANCOVAs, testing for heterogeneity of effects across studies, with study as the factor, post-treatment SR subscale score as the dependent variable, and pre-treatment SR subscale score as the covariate, indicated no significant study effect among the Fb group in ITT analysis, F(3, 261) = 0.72, p = 0.542, or efficacy analysis, F(3, 130) = 1.96, p = 0.124. The results showed a significant heterogeneity for the NOT TAU group, F(3, 307) = 3.73, p = 0.012. Effect sizes were calculated using both meta-analytic and mega-analytic approaches. Meta-analytic results yielded slightly smaller effect sizes than mega-analytic approach, but the differences were .02 and .04 for the ITT and efficacy analyses, respectively. Thus, only the mega-analytic results are presented. Mean pre-treatment SR subscale scores between NOT Fb and NOT TAU were equivalent in both ITT analysis, t(576) = 0.74, p = 0.463, and efficacy analysis, t(445) = -0.25, p = 0.805.

The combined effect size between the NOT Fb and NOT TAU was g = -0.24, p = 0.005, 95% CI [-0.40, -0.07], equivalent of 1.1 OQ point difference on average, favoring the NOT Fb group. Individual effect sizes ranged from g = -0.37, p = 0.005, 95% CI [-0.62, -0.11] to g = 0.08, p = 0.754, 95% CI [-0.42, 0.57]. In efficacy analysis the effect size was g = -0.48, p <0.001, 95% CI [-0.69, -0.28], equivalent of 2.3 OQ points difference, favoring the NOT Fb group. Individual effect sizes ranged from g = -0.65, p < 0.001, 95% CI [-0.98, -0.32] to g = -0.07, p = 0.801, 95% CI [-0.62, 0.48]. These results suggest that Fb has a small to moderate

advantage over TAU in improving outcome related to the social role performance of NOT patients.

Patient/therapist feedback effect. The results of ANCOVAs, testing the possible heterogeneity of effects did not reach statistical significance in either ITT analysis, $F(2, 217) = 2.42$, $p = 0.092$, or efficacy analysis, $F(2, 172) = 2.54$, $p = 0.082$. The between-group difference on pre-treatment SR subscale scores did not reach statistical significance in either ITT analysis, $t(405) = 0.97$, $p = 0.334$, or efficacy analysis, $t(321) = 0.26$, $p = 0.792$, indicating that the two groups were comparable in the level of disturbance at pre-treatment as measured by the SR subscale. The combined effect size for the difference in mean post-treatment SR subscale scores in ITT analysis was $g = -0.22$, $p = 0.137$, 95% CI [-0.52, 0.07], equivalent of 1.0 OQ point difference on average, favoring the NOT P/T Fb; however, this difference is not statistically significant. Individual effect sizes ranged from $g = -0.42$, $p = 0.008$, 95% CI [-0.73, -0.11] to $g = 0.02$, $p = 0.907$, 95% CI [-0.28, 0.31]. In efficacy analysis, the combined effect size was $g = -0.24$, $p = 0.193$, 95% CI [-0.61, 0.12] with individual effect sizes ranging from $g = -0.53$, $p = 0.002$, 95% CI [-0.86, -0.20] to $g = 0.01$, $p = 0.934$, 95% CI [-0.33, 0.35]. These results suggest that P/T Fb intervention adds little incremental benefit to the Fb. The ranges of effect sizes, however, suggest that some studies demonstrated more incremental benefits than others.

Prior to comparing the NOT P/T Fb group with NOT TAU using a mega-analytic approach, independent samples *t*-tests were conducted to compare the mean pre-treatment SR subscale scores to test for equivalence of the two groups at pre-treatment. The results did not reach significance in either ITT analysis, $t(531) = 0.583$, $p = 0.560$, or efficacy analysis, $t(486) = 0.002$, $p = 0.998$, indicating that both groups were equivalent at pre-treatment. The effect size for ITT analysis was $g = -0.40$, $p < 0.001$, 95% CI [-0.57, -0.23], equivalent of 1.9 OQ point

difference on average at post-treatment, favoring the NOT P/T Fb group. The effect size for efficacy analysis was g = -0.56, p < 0.001, 95% CI [-0.75, -0.37], equivalent of 2.5 OQ point difference on average at post-treatment.

Clinical Support Tools feedback (CST Fb) effects. Similar to the results of other subscale score analyses, the results of ANCOVAs testing for heterogeneity among the CST Fb group showed significant study effect in ITT analysis, $F(2, 408) = 4.09$, $p = 0.017$, but not in efficacy analysis, $F(2, 210) = 0.69$, $p = 0.504$. As in the case with other subscale analyses, the primary cause of heterogeneity in ITT analysis appears to be the greater degree of difference in pre-post change scores observed among patients in study 3. There was no significant difference in mean pre-treatment SR subscale scores between CST Fb and NOT Fb in either ITT analysis, $t(654) = 0.42$, $p = 0.677$, or efficacy analysis, $t(380) = 0.29$, $p = 0.776$, indicating that these groups were equivalent at pre-treatment in terms of social role performance as measured by the SR subscale scores.

The combined effect size, comparing the mean post-treatment SR scores between CST Fb and NOT Fb, was significant at .05 level in ITT analysis, g = -0.20, p = 0.018, 95% CI [-0.36, -0.03], equivalent of 0.9 OQ point difference on average, favoring the CST Fb group. Individual effect sizes ranged from g = -0.36, p = 0.036, 95% CI [-0.70, -0.02] to g = -0.08, p = 0.548, 95% CI [-0.35, 0.18]. The combined effect size in efficacy analysis was also significant, g = -0.28, p = 0.006, 95% CI [-0.49, -0.08], equivalent of 1.2 OQ point difference on average, with individual effect sizes ranging from g = -0.36, p = 0.027, 95% CI [-0.68, -0.04] to g = -0.15, p = 0.509, 95% CI [-0.59, 0.29]. These results suggest that CST Fb may add small incremental benefits over the Fb intervention in terms of social role performance at post-treatment.

The pooled data of the CST Fb was also compared with the pooled NOT TAU. No significant pre-treatment mean SR score differences was found in either ITT analysis, $t(722) = -0.26$, $p = 0.796$ or efficacy analysis, $t(524) = -0.49$, $p = 0.625$, indicating that both groups were equivalent in social role performance at pre-treatment as measured by the SR subscale. The effect size for ITT analysis was $g = -0.38$, $p < 0.001$, 95% CI [-0.53, -0.23], equivalent of 1.7 OQ point difference on average, favoring the CST Fb group. In efficacy analysis, $g = -0.64$, $p < 0.001$, 95% CI [-0.81, -0.46], equivalent of 2.9 OQ point difference on average.

Pre-post change in OQ subscale scores. The effect sizes for pre-treatment to post-treatment change scores in OQ subscales were calculated for each feedback condition, using mega-analytic approaches. Two types of effect sizes were calculated for pre-post change subscale scores. First pre-post change effect size was obtained and reported on the pooled dataset for each treatment group (See Table 15). The formulae for calculating pre-post change effect size (in Hedges's *g*) are presented in Appendix B. A mega-analytic approach was used to calculate the effect sizes in terms of between-group difference in the amount of mean pre-post change scores on the OQ subscales. The formulae for calculating the standardized difference in mean pre-post change scores are also presented in Appendix B.

Pre-post change in OQ subscale scores by feedback condition.

Treatment as usual (TAU). The pre-post change effect sizes of the NOT TAU group for the SD scale, IR scale, and SR scale were $g = -0.14$, $p = 0.019$, 95% CI [-0.25, -0.02]; $g = 0.12$, $p = 0.018$, 95% CI [0.02, 0.22]; and $g = 0.09$, $p = 0.128$, 95% CI [-0.03, 0.21], respectively. These results suggest that NOT TAU group as a whole experienced a small degree of, but statistically significant, improvement in symptom distress, while experiencing a statistically significant worsening in disturbances related to interpersonal relations and no change in the social role

performance. The summary of pre-post effect sizes for NOT TAU group and other NOT groups are presented in Table 16.

Feedback (Fb) effect. In ITT analysis, the respective pre-post change effect sizes of the NOT Fb group for the SD scale, IR scale, and SR scale were g = -0.30, $p < 0.001$, 95% CI [-0.39, -0.21]; g = -0.06, p = 0.163, 95% CI [-0.15, 0.03]; and g = -0.13, p = 0.008, 95% CI [-0.23, -0.03]. In efficacy analysis, the respective effect sizes for the SD, IR, and SR subscales were g = -0.49, $p < 0.001$, 95% CI [-0.61, -0.37]; g = -0.17, p = 0.003, 95% CI [-0.29, -0.06]; and g = -0.25, $p < 0.001$, 95% CI [-0.39, -0.12]. The effect sizes for the NOT Fb are summarized in Table 17. These results indicate that the greatest reduction of distress in the NOT Fb group occurred in symptom distress. The IR subscale had the smallest effect sizes, suggesting the least change from pre-treatment to post-treatment among patients receiving the Fb intervention.

When the pre-post change effect sizes of the NOT Fb group were compared with those of NOT TAU, between-group effect sizes (Hedges's g) in ITT analysis for the SD, IR, and SR subscales were -0.20 (p = 0.015), -0.18 (p = 0.029), and -0.28 ($p < 0.001$), respectively. Although these group differences were statistically significant, the average amount of change that occurred in the IR and SR subscales was quite small for the NOT Fb patients, with pre-post change effect sizes of -0.06 and -0.13, respectively, in the intent-to-treat (ITT) analysis. The between group differences were significant because the NOT TAU patients on average experienced increase in the disturbances measured by the IR and SR subscales, g = 0.12 and g = 0.09, respectively, suggesting worsening of the outcome on these dimensions.

Symptom distress as measured by the SD subscale showed the most positive change in both the NOT Fb group (g = -0.30) and the NOT TAU (g = -0.14). In efficacy analyses, the effect sizes for the SD, IR, and SR subscales were -0.47, -0.36, and -0.41, respectively (all $p <$

0.001). Details of the results are presented on Table 18. These results suggest that patients in the NOT Fb on average experienced a greater degree of improvement in all three major domains of outcome as measured by the OQ. The results further suggest that between-group differences in the pre-treatment to post-treatment change are unitary across the SD, IR, and SR subscales.

Patient/therapist feedback (P/T Fb) effects. In ITT analysis, the pre-post effect sizes of the NOT P/T group for the SD, IR, and SR subscales were g = -0.42, $p < 0.001$, 95% CI [-0.56, -0.27]; g = -0.16, p = 0.025, 95% CI [-0.30, -0.02]; and g = -0.37, $p < 0.001$, 95% CI [-0.51, -0.23], respectively. In efficacy analysis, the effect sizes for the SD, IR, and SR subscales were g = -0.56, $p < 0.001$, 95% CI [-0.74, -0.37]; g = -0.28, p = 0.001, 95% CI [-0.44, -0.11]; and g = -0.50, $p < 0.001$, 95% CI [-0.68, -0.32], respectively. These results suggest that pre-treatment to post-treatment change in the SR subscale among the NOT P/T Fb patients showed a similar degree of change in the positive direction as the SD subscale.

As in other meta- and mega-analytic results presented in this study, the NOT P/T Fb group was compared to the NOT Fb group and NOT TAU. In ITT analysis, between-group difference effect sizes (Hedges's g) in pre-post change for the SD, IR, and SR subscales were -0.10 (p = 0.317), -0.07 (p = 0.509), -0.29 (p = 0.004), respectively. In efficacy analysis, effect sizes for the SD, IR, and SR subscales were -0.08 (p = 0.494), -0.13 (p = 0.237), and -0.24 (p = 0.029). These results suggest that the P/T Fb intervention has a small effect on improving social role performance-related outcome measured by the SR subscale in comparison to provision of progress feedback to therapists alone in NOT patients.

When comparing the NOT P/T Fb group with NOT TAU, the effect sizes from ITT analysis for the SD, IR, and SR subscales were -0.27 (p = 0.002), -0.29 ($p < 0.001$), and -0.44 ($p < 0.001$), respectively. In efficacy analysis, effect sizes for the SD, IR, and SR subscale scores

were -0.39, -0.41 and -0.53 (all $p < 0.001$), respectively. It is interesting to note that the patients in the NOT P/T Fb group on average endorsed the most change in the social role performance aspect of outcome in comparison to both NOT Fb and NOT TAU. The possible meanings of these findings, although reliable, are unknown at this time.

Clinical support tools (CST) effects. Pre-post change effect sizes for the SD, IR, and SR subscale scores in ITT analysis were $g = -0.43$, $p < 0.001$, 95% CI [-0.53, -0.33]; $g = -0.23$, $p < 0.001$, 95% CI [-0.33, -0.13]; and $g = -0.29$, $p < 0.001$, 95% CI [-0.29, -0.40], respectively. In efficacy analyses, the respective subscale pre-post change effect sizes were $g = -0.76$, $p < 0.001$, 95% CI [-0.91, -0.60]; $g = -0.46$, $p < 0.001$, 95% CI [-0.61, -0.32]; and $g = -0.54$, $p < 0.001$, 95% CI [-0.70, -0.38], respectively. These results suggest that, consistent with other feedback interventions, patients in the CST Fb condition experienced the most change in their symptom distress. Interestingly, contrary to other subscales, there was a greater degree of change on the interpersonally oriented outcome measured by the IR subscale. This may be due to the fact that the CST intervention calls clinician's attention toward intervening in the patient's life outside of the therapeutic hour.

When the pre-post change scores in subscales were compared to the NOT Fb group, the effect sizes (Hedges's g) for the SD, IR, and SR subscale scores in ITT analysis were -0.13 ($p = 0.107$), -0.12 ($p = 0.137$), and -0.18 ($p = 0.026$), respectively. In efficacy analysis, the effect sizes for SD, IR, and SR subscales in comparison to NOT Fb were -0.23 ($p = 0.024$), -0.29 ($p = 0.005$), and -0.27 ($p = 0.009$), respectively. These results suggest that, in comparison to NOT Fb patients in the efficacy sample, the CST Fb group, who stayed in treatment long enough to reap the benefit of this treatment, experienced statistically significant pre-post improvement in all dimensions of outcome measured by the OQ.

When the subscale pre-post change scores of the CST Fb group were compared to those of the NOT TAU group, the effect sizes (Hedges's *g*) for the SD, IR, and SR subscale scores in ITT analysis were -0.30, -0.36, and -0.34 (all $p < 0.001$), respectively. In efficacy analysis, the effect sizes for SD, IR, and SR subscales in comparison to NOT TAU were -0.60, -0.60 and -0.57 (all $p < 0.001$), respectively. These results indicate that the CST Fb group as a whole experienced moderate pre-treatment to post-treatment change in all dimensions measured by the OQ.

Differences in clinical significance. As in the case with evaluating the OQ total scale scores based treatment outcome, treatment outcome at the subscale level was evaluated on the basis of the rates and odds of clinically significant change status at termination. The *n* and percentage of clinical significance classification of patient outcome at termination for each subscale were aggregated by treatment condition and presented in Tables 17 (SD scale), 18 (IR scale), and 19 (SR scale). These results indicate that the percentages of clinical significance classification for the SD subscale closely resembled those observed for the OQ total scale classification, with an exception of a lower rate of deterioration and an increased rate of clinically significant improvement among the NOT TAU group. Conversely, the rates of reliable deterioration and clinically significant improvement in all treatment conditions for the IR and SR subscales were considerably lower than those observed in the OQ total scale scores based classification results. Indeed, nearly 81 to 85% of NOT patients in all of the treatment conditions in ITT analysis were classified as "no change" cases in the domains of outcome measured by the IR and SR subscales. Even in efficacy analysis, approximately 76 % to 81% of NOT patients in all treatment conditions were classified as "no changers."

Various possible interpretations of these results seem plausible. Such results may suggest that the outcome related to one's interpersonal relations and social role performance tend to be more stable in terms of change than symptom distress. Another related view of these results may suggest the challenging nature of improving clients' relational concerns and social role performance through psychotherapy, which in many cases is a time-limited enterprise. Yet, another view may be offered in an acknowledgement of the magnitude of external and contextual influences on patients' relational experiences and social role performance. That is, in contrast to symptom distress-related outcome, which may be more amiable to, treatment outcome related to interpersonal relationships and social role performance may involve more external and contextual influences that are beyond the reach of therapeutic effects of psychotherapy in general. Thus, psychotherapy, at least the type of modality utilized in the studies reviewed in this review (i.e., individual therapy), may be limited in its outcome enhancing influences on interpersonal relationships and social role performance.

In the following sections, mega-analytic comparisons of the odds of deterioration and clinically significant improvement between various treatment groups are presented. The summaries of effect sizes, including 95% CI for each effect size are presented in Tables 20 and 21.

Feedback effect (Fb). When the odds of patient deterioration/reliable worsening at termination for the NOT Fb group were compared against those of the NOT TAU group, ITT analysis yielded the following odds ratios (*OR*) for the SD, IR, and SR subscales, respectively: 0.66 ($p = 0.094$), 0.93 ($p = 0.807$), and 0.64 ($p = 0.192$). In efficacy analysis, the *OR* for the SD, IR, and SR subscales were 0.43 ($p = 0.020$), 0.67 ($p = 0.313$), and 0.53 ($p = 0.179$), respectively. When the odds of patients achieving clinically significant improvement were compared between

the NOT Fb group and NOT TAU, the *OR* from the ITT analysis for the SD, IR, and SR subscales were 1.10 ($p = 0.595$), 1.45 ($p = 0.240$), and 2.16 ($p = 0.006$). In efficacy analysis, the *OR* for the SD, IR, and SR subscales were 1.72 ($p = 0.011$), 2.25 ($p = 0.018$), and 2.86 ($p < 0.001$), respectively. These results suggest that provision of progress feedback to patients helped prevent the occurrence of deterioration in symptom distress, especially for patients who stayed long enough to reap the benefit of this treatment. Although the odds of deterioration seemed lower for the NOT Fb group in other domains of outcome, the results did not reach statistical significance. The results further suggest that patients in the NOT Fb group had two to three times higher odds of achieving clinically significant improvement in social role performance than those in NOT TAU. Similar odds of clinically significant improvement were observed in the outcome related to interpersonal relations.

Patient/therapist feedback (P/T Fb) effect. When the odds of patient deterioration/reliable worsening at termination for the NOT P/T Fb group were compared against those of NOT Fb, the ITT analysis yielded the following odds ratios (*OR*) for the SD, IR, and SR subscales, respectively: 1.11 ($p = 0.718$), 0.77 ($p = 0.493$), and 0.64 ($p = 0.289$). In efficacy analysis, the *OR* for the SD, IR, and SR subscales were 1.40 ($p = 0.369$), 0.82 ($p = 0.662$), and 0.94 ($p = 0.895$), respectively. When the odds of patients achieving clinically significant improvement were compared between the NOT Fb group and NOT TAU, the *OR* from the ITT analysis for the SD, IR, and SR subscales were 1.34 ($p = 0.172$), 1.43 ($p = 0.245$), and 1.87 ($p = 0.053$), respectively. In efficacy analysis, the *OR* for the SD, IR, and SR subscales were, 1.28 ($p = 0.283$), 1.53 ($p = 0.183$), 1.96 ($p = 0.046$) respectively. These results suggest that provision of progress feedback to both patients and therapists in NOT cases had no significant incremental benefit in reducing deterioration at the subscale level. The odds of clinically significant

improvement among the NOT P/T Fb group were approximately two times higher for social role performance related outcome. It is interesting to note that the odds ratios for deterioration were higher in efficacy analysis than ITT analysis. A closer look at the differences in odds of patient deterioration, however, do not indicate higher odds of deterioration among patients in the efficacy sample than those in the ITT sample for the NOT P/T Fb group. The percentages of deterioration were slightly lower among the efficacy group than the ITT; however, the degree of reduction in the odds of deterioration among the NOT Fb group was greater than that of the NOT P/T Fb, which led to increase in the *OR*s.

The odds of deterioration and clinically significant improvement were also compared between the NOT P/T group and the NOT TAU group. In ITT analysis, when odds of patient deterioration were compared between the two groups, odds ratios for the SD, IR, and SR subscales were 0.84 ($p = 0.496$), 0.68 ($p = 0.250$), and 0.60 ($p = 0.173$), respectively. In efficacy analysis, odds ratios for the SD, IR, and SR subscales were 0.73 ($p = 0.255$), 0.63 ($p = 0.212$), and 0.62 ($p = 0.231$), respectively. When the odds of patients achieving clinically significant change were compared between the NOT P/T Fb and NOT TAU groups, odds ratios for the SD, IR, and SR subscales were 1.43 ($p = 0.054$), 2.38 ($p = 0.004$), and 2.21 ($p = 0.006$), respectively. In efficacy analysis, the odds ratios for the respective subscales were 1.80 ($p = 0.003$), 3.12 ($p = 0.001$), and 2.79 ($p < 0.001$). These results suggest that patients in the NOT P/T Fb group as a whole did not experience significant reduction in the odds of deterioration; however, the odds of clinically significant improvement was higher in all subscales, more notably in the outcomes related to interpersonal relations and social role performance.

Clinical support tools (CSTs) effects. The comparisons of odds of deterioration between the CST Fb group and the NOT Fb group in ITT analysis yielded the following effect sizes

(*OR*s) for the SD, IR, and SR subscales: 0.87 ($p = 0.594$), 0.65 ($p = 0.225$), and 0.51 ($p = 0.055$), respectively. In efficacy analysis, the odds ratios for the respective subscales were 0.77 ($p = 0.594$), 0.60 ($p = 0.317$), and 0.42 ($p = 0.129$). The comparisons of the odds of patients achieving clinically significant improvement between the CST Fb group and the NOT Fb group yielded the following odds ratios for the SD, IR, and SR subscales: 1.32 ($p = 0.110$), 1.43 ($p = 0.153$), and 1.67 ($p = 0.048$), respectively. In efficacy analysis, the odds ratios for the respective subscales were 1.69 ($p = 0.013$), 1.81 ($p = 0.039$), and 1.97 ($p = 0.022$). These results suggest that CST's incremental benefit in reducing the odds of the occurrence of deterioration in outcomes measured by the OQ subscale did not reach statistical significance. The incremental outcome enhancing benefit was found in the increased odds of patients achieving clinically significant improvement in all subscales among the efficacy sample.

When the odds of deterioration were compared between the CST Fb group and NOT TAU were compared in ITT analysis, odds ratios for the SD, IR, and SR subscales were 0.66 ($p = 0.059$), 0.43 ($p = 0.006$), and 0.46 ($p = 0.020$), respectively. In efficacy analysis, the odds ratios for the same comparisons for the respective subscales were 0.35 ($p = 0.001$), 0.32 ($p = 0.008$), and 0.27 ($p = 0.010$). Comparisons between the CST Fb and NOT TAU groups in the odds of patients achieving clinically significant improvement yielded the following odds ratios for the SD, IR, and SR subscales: 1.42 ($p = 0.030$), 2.49 ($p < 0.001$), and 2.18 ($p = 0.002$), respectively. In efficacy analysis, the same group comparisons resulted in the following odds ratios for the respective subscales: 2.52 ($p < 0.001$), 3.78 ($p < 0.001$), and 3.16 ($p < 0.001$). These results suggest that patients in CST Fb group had significantly lower odds of deterioration—approximately a half to a third—in all three domains of outcome measured by the OQ subscales. The odds of patients in the CST Fb group achieving clinically significant

improvement were significantly higher in all subscales, notably so on interpersonal relations and social role performance subscales.

Discussion on Study 2

In this study, the effects of various progress feedback interventions were evaluated at the OQ subscale level on various dimensions of treatment outcomes: differences in mean post-treatment scores across treatment groups, pre-treatment to post-treatment score change in each treatment condition, differences in pre-post change scores across treatment groups, odds of deterioration, and odds of achieving clinically significant improvement. The findings from these analyses revealed the effects of feedback interventions on more specific aspects of treatment outcome as measured by the OQ subscale scores, which were not available from the analyses based on the OQ total scale scores alone. To summarize the findings from this study in conjunction with the findings from the OQ total scale based analyses, the main findings for each feedback treatment are now discussed in turn.

Progress feedback to not-on-track patients (NOT Fb). When comparing the post-treatment OQ subscale scores, the average not-on-track patient whose therapists received progress feedback (NOT Fb) experienced superior outcome than 60% of NOT patients whose therapists received no progress feedback (NOT TAU) on all subscales (effect sizes ranging from g = -0.26 to g = -0.24). These results were comparable to the effect size for the same group comparison on the OQ total scale. When the same comparisons were made between the NOT Fb patients, who stayed in treatment long enough and completed at least a minimum number of the OQ to measure the effects of the feedback intervention, and the NOT TAU, the results likewise showed similarities between the effect sizes for the total scale (g = -0.53) and those for the Symptom Distress (SD) and Social role performance (SR) subscales (g = -0.52 and -0.48,

respectively). These results suggest that the average patient in the NOT Fb group in the efficacy sample experienced greater reduction of symptom distress and disturbances related to social role performance than approximately 70% of NOT TAU patients.

When the amount of change in the OQ subscale scores from the pre-treatment to post-treatment were compared, the effect sizes of the NOT Fb group in comparison to the NOT TAU group ranged from g = -0.28 for the SR subscale to g = -0.18 for the IR subscale. The degree of pre-treatment to post-treatment change increased in all subscales among NOT Fb patients in the efficacy sample, resulting in greater between-group difference in pre-post change (effect sizes ranging from g = -0.47 for the SD subscale to g = -0.36 for the IR subscale). These effect sizes were smaller than the effect size obtained from the difference in pre-post change on the total scale scores between NOT Fb and NOT TAU (g = -0.60). These results suggest that while the average patient in the NOT Fb experienced greater pre-post change on the total scale score than approximately 73% of patients in the NOT TAU group, the average patient's pre-post change on the subscale scores were superior than approximately 64 to 68% of the NOT TAU patients.

The analyses of clinical significance indicated that, the NOT Fb group had 12%, 9%, and 6% of deterioration rate on the SD, IR, and SR subscales, respectively, in ITT analysis, while the respective percentages reduced to 8%, 7%, and 5% in efficacy analysis. The rates of patients achieving clinically significant improvement in the SD, IR, and SR subscale scores in the ITT analyses were 31%, 10%, and 12%, respectively, while the respective percentages in the efficacy analyses increased to 39%, 13%, and 13%. These results were contrasted to deterioration rates among the NOT TAU in the SD, IR, and SR subscales (16%, 10%, and 8%, respectively) as well as improvement in the respective subscales, 29%, 6%, and 7%. When odds of deterioration were compared against those of the NOT TAU group, patients in the NOT Fb group had lower odds of

deterioration in the SD and SR subscales, although the SD subscale reached statistical significance in efficacy analysis. The odds of deterioration on the IR subscale for the NOT Fb group were near identical with those for the NOT TAU group in ITT analysis, but improved in the efficacy analysis. When the odds of clinically significant improvement were compared across groups, the odds were highest on the SR subscale in both ITT and efficacy analyses, favoring the NOT Fb group (*OR* = 2.2 and *OR* = 2.9, respectively). The SD subscale had the lowest odds ratios; however, these results need to be understood in the context that the IR and SR subscales generally had lower base rates of clinically significant achievement than the SD subscale. Indeed, despite the higher odds ratios on the IR and SR subscales, the sheer percentage of clinically significant improvement occurred in the SD subscale for all treatment groups surpassed those found on the IR and SR subscales.

Progress feedback to both not-on-track patients and clinicians (NOT P/T Fb). Provision of progress feedback to both patients and therapists was tested in previous feedback studies to investigate its potential incremental benefits in preventing client deterioration and enhancing outcome in comparison to provision of feedback to therapists alone. Overall, there were only a few aspects of outcomes in which P/T Fb seemed to have had positive effects on NOT patients. When comparing post-treatment OQ subscale scores of the NOT P/T Fb group with those of NOT Fb group, the results showed small effect sizes favoring the NOT P/T Fb group, but not at the statistically significant level. These results were consistent with the findings from the meta-analysis of the OQ total scale scores (see Study 1 or Shimokawa, Lambert, & Smart, 2010). The largest of the treatment effects (although still not statistically significant) were found in the SR subscale. In terms of the amount of pre-treatment to post-treatment change scores, the NOT P/T Fb group showed very small effects on the SD and IR

subscales. Small, but statistically significant effects were found on the SR subscale, favoring the NOT P/T Fb group over the NOT Fb group in both ITT and efficacy analyses (g = -0.29 and g = -0.24). When the odds of patient deterioration were compared to the NOT Fb group, the NOT P/T Fb group had higher odds of deterioration on the SD subscale in both ITT and efficacy analyses, *OR* = 1.11 and *OR* = 1.40, respectively, although these differences were not statistically significant. These increased odds of deterioration were similar to the trend found on the OQ total scale analyses. Such a trend was not found on the IR and SR subscales. The subscale analyses were, thus, helpful in identifying what aspects of the outcome (i.e., symptom distress) that contributed to the increased odds of deterioration observed in the total scale score analyses. The reasons for potentially higher likelihood of deterioration in symptom distress are unknown at this time. The odds of patients achieving clinically significant improvement were higher for the NOT P/T Fb group than the NOT Fb group in all domains of outcome measured by the subscales, although the results reached statistical significance only on the SR subscale. It is interesting to note these significantly greater effects found on the SR subscale among the NOT P/T Fb group. It appears that the potential "polarizing effect" of the P/T Fb among not-on-track patients (Shimokawa, Lambert, & Smart, 2010) may have occurred most notably in patient symptom distress in terms of deterioration and social role performance in terms of clinically significant improvement.

Clinical support tools and progress feedback to not-on-track patients (CST Fb). Potential incremental benefits of clinical support tools on patient outcome at the OQ subscale level were evaluated against the outcome of provision of progress feedback alone (NOT Fb). The post-test subscale score comparisons between the CST Fb and NOT Fb groups yielded small effects in both ITT and efficacy analysis, with the SR subscale being the only subscale in which

statistically significant effect sizes were observed in both ITT analysis, g = -0.20, and efficacy analysis, g = -0.28. When the amount of pre-treatment to post-treatment change in subscales was examined mega-analytically, the CST Fb group resulted in the greatest degree of change on all of the subscales in the efficacy analysis, although the NOT P/T Fb group yielded a greater effect size on the SR subscale in ITT analysis. Between-group comparisons of pre-post change scores on the OQ subscales suggest that the CST Fb group was somewhat more helpful than the NOT Fb group in patients experiencing positive change on all three dimensions of the outcome (g = -0.23 on the SD subscale; g = -0.29 on the IR subscale; and g = -0.27 on the SR subscale). These results indicate that the average patient in the CST Fb group, who met the minimum criteria to be included in the efficacy analysis, had better outcome than approximately 59 to 62% of patients in the NOT Fb group on the basis of pre-post changes in subscales. The results also suggest that the average patient in the CST Fb group fared better than approximately 73% of patients who did not receive any feedback treatments. While the SR subscale scores of the CST Fb group resulted in an effect size similar to those of the NOT P/T Fb group, the IR subscale resulted in better outcome favoring the CST Fb group. When the odds of deterioration were compared with those of the NOT Fb group, the CST Fb group appeared to reduce the odds of deterioration by 0.9 to 0.4 times, but not at the statistically significant level. In comparison to the NOT Fb group, the odds of patients achieving clinically significant improvement on the SR subscale in the ITT analysis and on all of the subscales in the efficacy analyses were significant. These results suggest that the odds of CST Fb patients achieving clinically significant improvement on dimensions measured by the OQ subscales are 1.7 to 2.0 times higher than those in the Fb group. In comparison to the NOT TAU group, these odds increase to 2.2 to 3.8 times higher. Based on the aforementioned results, the CST Fb appears to have incremental benefits in

enhancing outcomes related to symptom distress, interpersonal relationships, and social role performance. The incremental benefit of the CST Fb intervention on the IR subscale may suggest the intervention effects that were unique to the nature of the CST Fb intervention. In particular, the measures included as part of the Clinical Support Tools assessed the not-on-track patients' social support. Given the similarity between the construct of social support and the interpersonal relationships, some of the greater effects found on the IR subscale scores among patients in the CST Fb group may have been the result of the Clinical Support Tools being utilized in the clinical work.

Limitations of study 2. All of the limitations discussed in study 1 applied to this study because this study utilized the same methodologies as study 1 (except for the mega-analytic approach used on the analyses of group differences in pre-post change scores in this study). Some additional limitations should be noted that were unique to this study. Although the accuracy of the degree of data match was deemed adequate in this study in favor of conserving data, to the extent the discrepancy between the OQ scores in the original datasets and the merged datasets occurred, the presence of error was likely, especially in a very small minority of cases where the discrepancy was large. Conducting statistical analyses only on cases with reasonable accuracy in the match may reduce the "noise" introduced by those observations with discrepant scores in the match. Even then, however, the results of the analyses are expected to be similar, given the small number of cases with such discrepantly matched scores.

Another limitation of this study and of this line of research was the exclusive use of OQ total score in outcome monitoring and feedback provision. Although extensive examination of OQ subscales was performed in this study, the original feedback studies did not utilize the subscale information in provision of progress feedback. Thus, the treatment effects reflected at

the subscale level were still effects based on the OQ total scale score system. As already discussed earlier, subscale based information could provide unique clinical information that could not be captured by the total scale score alone. One practical advantage of subscale-based feedback system is that it requires no more burdens on patients other than filling out the OQ as they normally do. Thus, the future development and implementation of the OQ subscale score based progress feedback, in addition to the well-established total scale score based system may further enhance the clinical utility of this quality assurance system.

Study 3: Multi-level Modeling of Change in Patient Outcome

Method

Participants and procedures. Given the emphasis of the psychotherapy quality assurance system under study, only patients who were identified as signal alarm cases (not-on-track or NOT; n = 1382) were analyzed in this study. To be consistent with the meta-analytic and mega-analytic portion of this series of quantitative reviews, the data of NOT patients were analyzed using the same inclusion criteria, thus obtaining the estimates for both intent-to-treatment (ITT) sample and efficacy sample. It should be noted that these inclusion criteria pose some important limitations in multilevel analyses. Specifically, in meta- and mega-analyses of the ITT samples, the last observation carried forward method was used in cases where the post-test score was missing. In case of NOT patients who left treatment after the first warning event, the OQ score at the time of warning was also treated as the post-treatment score. This procedure was used to obtain a conservative estimate of the effects of treatment on patients. As long as this estimate is accepted by the researcher as a reasonable estimate of the post-treatment score, research questions regarding the *amount* of change may not pose a serious issue. In the case of multilevel analyses, however, the research question frequently involves change, which occurs as

a function of time. Thus, each data point connotes both change in terms of a measurement unit and a temporal time associated with it. Because of this, if one treats a given score as a substitution of another score at a different time point (e.g., an OQ score at the time of warning signal that is also a last observation of a given patient as the patient's post-treatment score), such a procedure automatically produces a slope of zero for that patient. Because of this, the ITT analyses with the inclusion of score substitution process in patients with missing scores may introduce bias in the estimates, especially in the slope. Thus, another criterion was applied to both ITT and efficacy analyses to examine the statistical models with no substitution of scores. By applying this criterion, only the actual data points obtained through the courses of studies were included. Patients without an OQ score after the first signal alarm event, for instance, contributed only to the estimation of the intercept at the first warning event, but not to the slope. Addition of this criterion doubled the number of analyses to be performed[11]. Thus, for practical reasons, not all of the results from the analyses based on alternative inclusion criteria are reported.

Statistical analysis. In studies examining individual change, especially in studies involving the use of multiple-time-point designs (such as the studies included in this study), multi-level analyses are more appropriate than studying only the pre- and post-treatment data. Multi-level analyses are called by various names, including the random-effects model, the general mixed-linear model, and the hierarchical linear model (Raudenbush & Bryk, 2002). Multilevel modeling of change allows modeling of patients within the individual as well as the

[11] In addition to ITT and efficacy analysis, separate analyses were conducted with and without 355 cases that belonged to patients who participated in studies more than once. All of the combinations of these analyses were conducted separately with and without observations with substitution of scores as described. Given that repeated participants comprised of about 5% of all participants (N = 6,151) and that results were comparable, this study presents only the results based on repeated participants included because inclusion of repeated participants in treatment is more reflective of typical clinical practice.

between-individual and group levels. Thus, the use of multilevel analyses provides the information that was not available in the traditional meta-analysis, which primarily examines the differences in group sample means. In this study, the linear mixed models function of SPSS Advanced Models 15.0 for Windows was used. Full maximum likelihood estimation method was used.

Alternative models of change. To determine the appropriate level-1 sub-model of change, empirical growth plots of NOT patients with person specific regression lines were inspected. This initial inspection presented with a wide variety of individual change patterns. In essence, by definition NOT patients deviated from the expected recovery curves defined by the OQ algorithms that were based on the dose-response models of change. Although signal warnings were operationally defined by the algorithms for predicting treatment failure as OQ points of marked deviation (i.e., elevation or lack of expected reduction in OQ points) from expected recovery pattern, such elevation occurred in various ways. For some patients, signal warning appeared to have occurred as an outlier "peak" event, while for some individuals the event of first warning seemed to mark the beginning of a phase with higher disturbance. The timing and number of, and space between, warning events also varied across individuals, some occurring close to each other whereas others occurring at what appeared to be separate phases of worsening. For a majority of patients, the first signal warning seemed to have occurred after a few sessions into treatment. For NOT patients who stayed in treatment for shorter periods of time, the change in OQ scores appeared to be linear, especially those who did not return to treatment after experiencing elevation in disturbance. For those who stayed in treatment longer, the change appeared to occur more non-linearly than linearly, presenting multiple modes of

worsening and improving. Describing individual empirical growth plots in a systematic manner seems daunting if not impossible as the patterns of change appeared quite diverse.

In the initial stage of model building, following the recommendations by Singer and Willet (2003), the unconditional means model and unconditional growth models were fit on the data of patients who were classified as not-on-track (NOT; n = 1382). These initial models were used to obtain the "baseline" of within-person and between-person variability to assess the incremental fit of, and the variance of change parameters explained by, more complex models. Finding non-linear growth pattern at the leve-1 submodel of change that applies *across* individuals can be a very tedious process (Singer & Willet, 2003). Various patterns of level-1 submodels of change, including the following were initially examined: linear model of change from intake to termination, piecewise linear model of change (Gallop & Tasca, 2009) with an added slope from the time of first signal warning feedback, log linear model of change from intake to termination, linear models of change with discontinuity in slope at the first signal warning, and a combination of linear slope and discontinuity in intercept and slope at the time of first warning.

To fit the patterns of change in linear regression models, time was treated in several ways at the initial stage of model building. First the session number without any transformation was used. Based on the findings from dose-response research (e.g., Lutz, Lowry, Kopta, Einstein, & Howard, 2001; Lutz, Martinovich, & Howard, 1999), natural logarithm of time (i.e., session number + 1) was also used. However, this study's exclusive focus on patients at risk of treatment failure suggested that patterns of change among a minority subset of patients may not fit widely recognized patterns of change reported in psychotherapy literature. For instance, by the definition of the not-on-track (NOT) classification of patient treatment progress, it was deemed

necessary that the level-1 submodel of change addressed the patterns of change before and after the first signal warning event. Combinations of separate pre and post-warning slopes were tested. Level-1 submodels with discontinuity in slope modeled the "shift" in linear slope after the first warning signal. The piecewise model assumes the presence of separate slopes for before and after a common time point where a presumed shift occurs. Although group mean comparison of OQ total scales at pre-test, the time of signal warning, and termination were different (e.g., Figure 2 of Harmon, et al., 2007), it was not clear if the pattern of change at the individual level was represented by separate slopes for before and after the first signal event as in the piecewise model. Although it was quite likely that some patients presented such a pattern, having a worsening trajectory from the beginning of treatment to the first warning followed by a change in the direction of recovery, fitting this model to all not on track patients presented some questions at the theoretical level. Particularly, while a large body of psychotherapy research findings demonstrate overall patterns of recovery among patients (whether it's dose response or good enough effect), is it reasonable to assume that not on track patients, as a group, for whatever reasons followed a worsening pattern from the beginning of treatment until they were identified as such? Undoubtedly some patients may enter treatment with already worsening trajectory in symptoms and functioning; however, is it reasonable to assume that this would be the norm for not-on-track patients as a group? Or, would the worsening be better construed as deviation from an expected recovery trajectory? The latter fits better with the guiding rationale for developing the OQ signal alarm system in the first place and seemed to fit the phenomenon of worsening in treatment better for NOT patients as a group.

The model of deviation from the expected trajectory seemed to fit the notion of discontinuity in intercept and slope due to fundamental shift in change process as discussed by

Singer and Willett (2003, chapter 5). This model reflects an underlying change pattern with a shifting point in disturbance level and different change trajectory after the first signal warning event. These two competing models were empirically tested. Each of the various models of change with varying level-1 submodels of change were initially tested without substantive predictors at level-2, except for a predictor indicating the number of sessions attended by patients as discussed in the next paragraph. This was conducted in order to select a model that seemed to better reflect the pattern of recovery at the individual level.

An initial primary research question of this study was whether differences existed across treatment groups in the rate of patient recovery. As discussed in the meta- and mega-analytic portion of this quantitative review, however, patients in various feedback conditions experienced superior outcome, but also stayed in treatment longer than the control group, especially *after* the first warning event. This finding suggests that feedback interventions might have had a retention effect, which in turn might have lead to improved outcome. If this potential retention effect explained most of the variance contributing to improved outcome in terms of the amount of change, differences in the rate of change would be unlikely to be found. Thus, the question in this study was to investigate how much, if any, difference existed in individual patients' rate of change that were explained by feedback treatments. If the rate of change differed across treatment groups, the amount of change found through the meta- and mega-analyses would be explained by both the retention effect and differential rates of recovery.

Issues of taking into account varying treatment duration among patients. One methodological issue regarding treating time needed to be addressed in this study. Because patients left treatment after varying number of sessions at will, there was no arbitrary number of sessions that could serve as a reference in estimating the "average slope." Although a number of

multilevel methods have been developed to handle "virtually every missing data problem" (Duncan, Duncan, & Strycker, 2006, p.179), the nature of varied treatment lengths in psychotherapy should not be construed as a missing data problem (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen., 2009). For example, when comparing the outcomes of clients who attended five sessions of therapy with those who attended 10 sessions of therapy, it is inappropriate to assume that the difference in session length (five sessions) is due to "missing sessions" because the termination of therapy was considered a naturally occurring end point often negotiated between therapists and patients.

As Baldwin and colleagues demonstrated differential rates of change as a function of the number of sessions patients attended, in this study the total number of sessions attended by each patient was tested as a level-2 predictor in each of the above submodels. Because of the markedly positively skewed distribution of total number of sessions attended by NOT patients, natural log transformation of number of sessions centered on the NOT patient mean was used to assess whether the rate of change after the first signal warning differed as a function of the total number of sessions patients attended. Centering of the number of sessions was performed because this variable's zero value was theoretically impossible (i.e., a patient attending zero session of therapy) and because of the intuitive interpretation of centering this predictor to the NOT sample mean (i.e., average number of sessions attended by NOT patients).

Based on the similar notion of heterogeneous recovery rates as a function of session attendance, a level-2 predictor of the total number of sessions *after* the first warning event was also tested. Initial testing of these level-2 predictors suggested a better model fit of stratification based on the number of sessions attended after the first warning than stratification based on the total number of session attendance.

Separate multilevel analyses were also conducted for OQ total scale scores and subscale scores for stratified bands of the number of sessions attended by NOT patients following the first signal warning event marking the 25th, 50th, 75th, 85th, 95th percentiles. Through this method, patients were classified into one of the following five groups based on the number of sessions attended after the first warning event: 2 sessions or less (n = 390, 28.2%), 3 to 4 sessions (n = 289, 20.9%), 5 to 8 sessions (n = 341, 24.7%), 9 to 11 sessions (n = 146, 10.6%), and 12 to 18 sessions (n = 147, 10.6%). Patients who attended beyond 18 sessions (above 95th percentile, n = 69, 5%) were not tested because of the smaller number of samples, making the models unreliable. Although a substantial number of patients attended only one or two sessions after the first signal warning event, multilevel models were not fitted for this group because data points per patient were too few. Because of the positively skewed distribution of session attendance, natural log of session number following the first signal warning was used. The model with level-2 predictors were estimated for each of the session bands.

Predictors. In level 2, variables that are commonly reported in this line of research as non-significant (e.g., age, gender, and diagnoses of patients) were not tested. Level-2 predictors included experimental conditions to which individuals were assigned, time variable (either linear or log transformed time variable), and length of treatment (i.e., number of sessions attended). Information such as the total number of session attended cannot be obtained prior to termination of treatment. Thus, the models presented in this study are exploratory and explanatory rather than predictive, that is, based on variables already known at the time of pre-treatment.

In the original feedback studies, with an exception of study 3, therapists served as the blocking variable to randomly assign clients to treatment conditions to control for therapist effects. Because of this design, experimental conditions essentially varied *within* therapists.

Even in study 3, although random assignment was not implemented, each therapist was given clients in both conditions based on the semester in which clients participated in treatment, thus making the experimental conditions crossed with therapists.

Variables that differ across therapists were initially planned to be analyzed at level 3. However, in the original feedback studies, datasets kept track of only the identification variable of one primary therapist per patient. The counseling center in which most of the feedback studies were conducted routinely assigned patients to different therapists after an intake sessions. In multilevel analyses of therapist's effects, each data point needed to be matched with the treating therapist for that session. Thus, change of therapists presented cross-classification of data. Such complex data structure cannot be handled with SPSS. Furthermore, very few, and no consistent, therapist variables were recorded in the original studies, which would make it difficult to study therapist effects. For these reasons therapist effects were not evaluated in this study.

Results

Descriptive statistics. NOT patients attended an average of 10.06 sessions (*SD* = 7.03, range = 2 – 60, *Mdn* = 8) and, after the first signal warning event, remained in treatment for an average of 6.58 sessions (*SD* = 6.36, range = 1 – 55, *Mdn* = 5). The mean OQ total scale score at intake was 80.16 (*SD* = 19.84, range = 18 – 149, *Mdn* = 79.00). The mean OQ score at the time of first signal warning was 89.01 (*SD* = 15.66, range = 58, *Mdn* = 87). The average OQ score at termination (the last recorded OQ for each patient) was 74.09 (*SD* = 22.55, range = 8 – 166, *Mdn* = 74.00). These statistics, especially the proximity of means and medians of OQ scores at three time points suggest that the distribution of OQ scores was normally distributed. For descriptive purposes, and to evaluate the possibility that session attendance after the first signal event

differed across treatment groups, *n* and percentage of the number of treatments NOT patients attended after the first signal are presented in Table 22 for each treatment group.

Rate of change from intake to termination. As noted earlier, prior to testing the effects of feedback interventions, models with various level-1 submodels of change and level-2 time predictors were tested. These analyses were conducted to gain broader understanding about how not-on-track patients change overtime. Patterns of change in patients were modeled for the entire duration of treatment. Analyses were conducted for the ITT and efficacy samples and the ITT sample without using substitution of scores, all pooled across treatment groups. To facilitate integrated understanding between the OQ total scale scores and subscale scores, the results of the total scale analyses and subscale analyses are presented and discussed in this study. The results of comparing some of the models of change tested based on the OQ total scale scores are presented in Table 23 (ITT sample), Table 24 (efficacy sample), Table 25 (ITT sample without last observation carried forward), Table 26 (efficacy sample without last observation carried forward). These models were based on linear time of change (i.e., session number and session number after first warning event). Table 27 (ITT sample), Table 28 (efficacy sample), Table 29 (ITT sample without last observation carried forward), and Table 30 present results of models based on natural log-transformation of time due to the possibility that the log linear pattern of change as discussed in the dose-response effect literature might better capture the pattern of change.

Model A in Tables 23 to 30 represent unconditional means model to partition the variability in OQ scores *within* and *across* individuals. Interclass correlation coefficients of 0.301 in the ITT sample, 0.258 in the efficacy sample, 0.295 in the ITT sample without the use of LOCF, and 0.240 in the efficacy sample without the use of LOCF indicate that approximately

24% to 30% of variability in OQ total scores, without considering the rate of change, is associated with *between* patients. Model B.1 in various tables represents the unconditional growth models which provide the baseline for each patient's variability around his or her true linear change trajectory (rate of change) as well as between patient variability in the change trajectory. Model C.1 stratified the number of sessions NOT patients attended by estimating the effects of the total number of sessions patients attended in addition to the function of linear time. Model D.1 included an additional slope parameter at the level-1 submodel of change, indicating the added or altered rate of change after the first warning signal event in addition to the overall linear rate of change. Model E.1 introduced another time-variant dichotomous predictor (labeled "Signal" on the tables) at level-1 indicating whether or not a given data point belonged to before or after the first signal warning event. This variable estimated the shift in intercept at the time of first warning in comparison to the intake. As discussed earlier, Model E.1 was the theoretical competitor to the piecewise model.

When the piecewise model and the discontinuous slope and intercept model (Model E.1) were compared, Model E.1 indicated a significantly better fit, $\chi^2(4) = 719.2$ to 1159.7 (depending on the inclusion criteria used), $p < 0.01$, favoring the discontinuous slope and intercept model. Model F.1 added a level-2 predictor (labeled #SessionPW) to Model E.1 to stratify the rate of change by the number of sessions attended after the first onset of a signal warning.

Model B.2 was the unconditional growth model based on the natural log transformation of session (time variable)—essentially a log transformed version of Model B.1. Comparison of Model B.1 and Model B.2 showed that, regardless of inclusion criteria applied, Model B.1 showed superior model fit, -2 log likelihood differences of 317.9 to 428.1, favoring Model B.1. *These results indicate that when modeling the entire course of treatment for NOT patients,*

modeling the rate of change on linear time (non transformed session number) is superior to natural log transformed time. Stratification of time as in Models C.1 and C.2 still favored a linear model of change. Because of this, subsequent discontinuous slope and intercept terms were added to an overall linear model of change. Model D.1 introduced a new linear slope from the time of first signal. Model D.2 instead added a new log transformed linear slope from the first signal warning. The comparison between Models D.1 and D.2 showed that the model fit was better for Model D.2 by -2 log likelihood differences of 91.4 to 153, *suggesting that after the onset of first signal warning, patients pattern of change may be better captured by log transformed time than the linear time*. Addition of elevation in both the intercept and slope in Models E.1 and E.2 similarly supported this conclusion. Models F.1 and F.2 stratified the rate of change after the first signal warning. In model F.1 and F.2, inclusion of random effect terms for both the linear time (session number) and the term indicating whether a given observation came from before or after the first warning ("Signal") did not achieve conversion. Removal of one or the other reached model conversion. Although statistically significant fixed and random effects existed with the linear time ("Session"), the random effects associated with this variable had a relatively small variability across individuals, suggesting that the linear rate of change from intake to termination had statistically significant variability among NOT patients, but this variability was relatively small in comparison to the sum of all the between-patient variability. Thus, then random effect term associated with the linear time was dropped. Furthermore, in favor of parsimony, fixed effects that were non-significant or significantly unimportant were not included in the model.

While it was possible that more complex non-linear patterns of change fitted the model better, addition of quadratic or cubic terms after the first warning could not be performed

because the computational demands exceeded the computer's capability. Thus, in this study Model F.2 was accepted as the best fitting model for the entire duration of treatment.

These results suggest that NOT patients as a group tend to have a slow rate of progress prior to being recognized as at risk for treatment failure. The recovery after the first warning is better captured by a log linear slope than a linear slope, suggesting that progress occur rapidly shortly after the first warning event rather than following a linear trajectory. The results also showed that patients who stay in treatment longer have slower rates of change, similar to the findings reported by Baldwin, et al. (2009) who found differential rates of change when modeling the entire course of treatment of a sample of a general counseling center patient population. At this point, it is important to address the consequences of using the last observation carried forward method in multilevel analyses of change. Comparison between models with LOCF and those without appear to suggest that the use of LOCF method affected the results to some degree, most notably the rates of change. Further analyses were thus conducted on ITT and efficacy samples without the use of LOCF method, as these sets of inclusion criteria were deemed most representative of the actually observed data.

Even though the aforementioned models of change provide an overview of how patients change based on the OQ total scale scores, it was not clear prior to investigation whether the OQ subscale scores followed the same trajectory. Similar procedures were followed for the OQ subscales analyses; however, given the number of analyses involved, and for practical reasons, only the final models (Models F.1 and F.2) are reported in Tables 31, 32, and 33, for SD, IR, and SR subscales, respectively. Given the number of models and sets of samples modeled, conducting model building to find the "best fitting model" for each subscale per each inclusion criteria seemed impractical. Thus, the data was submitted to the same models developed for the

OQ total scale score analyses to evaluate the fit within the parameters set by the models (with the exception of Models F.1 and F.2 where the models converged with an inclusion of linear trajectory across the entire course of treatment). The two models to which all subscale scores were fitted had only one difference—either linear time or log transformed time was used after the first signal warning, with all other parameters remaining the same across models. Model fit was superior for the log transformed model, suggesting that, just as on the OQ total scale scores, patients' recovery patterns at the subscale level followed closer to the log transformed time than the linear time.

Comparisons of rate of change across feedback treatment groups among NOT patients.

OQ total scale. To address the main question of this study (i.e., whether or not differences in the rate of recovery exist across feedback treatment groups), both linear and log linear models were fitted from the time of first signal warning to termination of treatment. The results of ITT multilevel analyses of patient change on the OQ total scale scores are presented in Tables 34 and 35, comparing the results of modeling of change by feedback treatment conditions based on non-transformed linear time (session after first episode of signal warning) and log transformed time.

The results of the efficacy analyses are presented in Tables 36 and 37. The ITT analyses suggest that when compared to NOT TAU patients with an average number of session attendance after the first episode of signal warning, patients with the same number of sessions in CST Fb and NOT Fb experience statistically significantly faster rate of recovery on the OQ total scores. For the CST Fb group, this translates to -3.57 OQ points per session average rate of change on the linear model of change and 10.56 OQ points per one log unit of session on average with

initially rapid reduction of scores followed by a more gradual rate of change (i.e., 10.56 OQ points reduction for the first 2.7 sessions and the same amount of reduction in the next 4.7 sessions span). For the NOT Fb group, the reduction in OQ points translates to 3.42 OQ points on average per session after first warning and 10.52 OQ points per one log unit of number of session on average after the first signal warning.

In separate analyses of rates of change grouped by differing number of sessions patients attended, however, results did not reach statistical significance in most bands of treatment lengths, except for the CST Fb patients among those who attended 12 to 18 sessions. In efficacy analyses, the rate of change were more similarly significant for the CST Fb patients attending the average number of sessions after first signal warning event, when compared to their NOT TAU counterparts. Based on the linear model of change, this rate of change translates to 3.61 OQ points on average per session. On the model based on log unit of time, this translates to about 11.56 OQ point reduction in the first 2.7 sessions and another 11.56 OQ points over the next 4.7 session span. In the separate analyses of rate of change based on varying session attendance, the rate of change was significant for the CST Fb group only in the 12 to 18 session span after the first signal warning event. However, statistically non-significant findings may have been due to reduction of power resulted by conducting separate analyses on smaller subsets of patients grouped by the number of session attendance. These combined results from both the ITT and efficacy analyses suggest that CST Fb and NOT Fb present with similar rates of change on the OQ total scale score that are significantly faster than the rate of recovery among those who received no feedback intervention.

Symptom distress (SD) subscale. The results of ITT analyses of patient change on the OQ Symptom Distress (SD) subscale score are presented in Tables 38 and 39, comparing the

results of models based on linear time and natural log transformation of time. The results of efficacy analyses are presented in Tables 40 and 41. As the difference of 164.8 in -2 log likelihood with the same number of parameters, the model based on log transformed model fit the data better. The linear model shrinks the variance in rate of change across individuals. The results also show that patients with the average number of session attendance after the first signal warning session in CST Fb group improve more rapidly in symptom reduction than those in NOT TAU attending the same number of sessions in both models. The recovery rate of those in NOT P/T Fb and NOT Fb are not expected to be significantly superior to those in TAU. The results of analyses conducted on separate bands of number of sessions attended by patients showed that the superior rate of change in symptom reduction for individuals in the CST Fb group occurs most saliently among patients who attended 12 to 18 sessions after the first onset of signal warning compared to those in the NOT TAU group (non-standardized regression coefficient of -4.26 per 1 log unit of session, $p = 0.006$ in efficacy analysis and non-standardized regression coefficient of -3.74, $p = 0.01$ in ITT analysis). Statistical difference in the rate of change on the SD subscale scores was not observed in other bands of session lengths. This non-significant finding, however, may have been due to the lack of statistical power.

Interpersonal relations (IR) subscale. The results of analyses of Interpersonal Relations (IR) subscale suggest that across all treatment groups, the rate of change after the first onset of signal warning is quite slow (see Tables 42 and 43 for ITT analyses and Tables 44 and 45 for efficacy analyses). However, when compared to NOT TAU patients attending the average number of sessions, patients in CST Fb group attending the same number of sessions experience statistically significant difference in the rate of change in ITT analysis when change was modeled in natural log transformation of time or linear time. Under the linear time of change

model, patients in NOT Fb may also experience statistically significant greater recovery rate in the IR subscale scores. As with the SD subscale analyses, NOT patients who stay in treatment longer are expected to have a slower rate of recovery as measured by the IR subscale scores (1.24 OQ points per 1 log unit increase). Separate analyses of rate of change suggest that CST Fb patients attending 9 to 11 sessions of treatment experience may experience a superior rate of change than their NOT TAU counterparts attending the same number of sessions (non-standardized regression coefficient of -0.98, $p = 0.04$). In efficacy analysis, the results were equivalent to those found in the ITT analysis, except patients in CST Fb who attend 12 to 18 sessions may also experience a statistically greater rate of change in the IR subscale scores.

Social role performance (SR) subscale. The results of analyses of rates of change on the SR subscale scores are presented in Tables 46 and 47 (ITT analyses) and Tables 48 and 49 (Efficacy analyses). The results suggest that addition of level-2 predictors (i.e., treatment group assignment and number of sessions attended) improved overall model fit at a statistically significant level in the -2 log likelihood (χ^2 difference of 42.9, $df = 10$, $p < 0.001$ in the log linear model) compared to unconditional growth model in efficacy analysis, other model fit indices mixed results in terms of statistically significant improvement. Between-individual differences in the rate of change, in particular, was quite small for both the linear and log linear models (more notably in the linear model), suggesting that both models may not be capturing the rate of change as they occur among patients. Another competing, and perhaps quite plausible, explanation may be that recovery in social role performance and interpersonal relations is slower compared to symptom reduction as reflected in the SD subscale scores. The results of ITT analyses showed that the rate of change among patients in NOT P/T Fb group was statistically significant at the .05 level. Statistically significant results were not found when the rate of

change was tested in separate analyses with patients grouped by the number of sessions they attended after first signal warning. No other significant findings were found in terms of differences in the rate of change on the SR subscale scores.

Discussion on Study 3

This was the first attempt to evaluate the OQ-based quality assurance system using multilevel analytic techniques. Initial investigation of the patterns of change suggested that NOT patients varied considerably in their patterns of change. Yet, despite such variability, some patterns of change fit better than others. The results of initial model fitting suggested that patients who are identified as at risk of treatment failure have a slow linear recovery trajectory that becomes disrupted at the time of first signal warning. At the time of first signal warning, there is an elevation of about 11 to 12 OQ points on average, though there is between-patient variability in the degree of elevation. After the point of first signal warning, patients' pattern of recovery is better captured by a non-linear (natural log of session number) trend with rapid initial recovery, which gradually slows down as patients stay in treatment longer. This general pattern of recovery was reflected in all subscales, though the rate of recovery was more gradual in the Interpersonal Relations subscale and the Social Role performance subscale than that in the Symptom Distress subscale.

At the subscale level, Patients in the CST Fb group appeared to have experienced a higher rate of recovery in the dimensions of outcome measured by the SD and IR subscales than those in TAU. The amount of effect as measured by the IR subscale was also shown to be significantly greater than those in TAU according to the meta-and mega-analysis portion of this study.

The primary research question in this study was: Is there a difference in the rate of change existed among treatment groups and control? The analyses of the OQ total scale score found statistically significant difference between the CST Fb group and NOT TAU, favoring the CST Fb group. It was unclear if this superior rate of change was present uniformly across varying numbers of sessions attended after the first signal warning. As reported earlier, there appeared to be between-group differences in the number of sessions patients attend after the first signal warning based on treatment assignment. Considering the findings from the meta- and mega-analyses demonstrating superior outcome experienced by patients in the CST Fb group, the prolonged treatment participation, and statistically significant faster rates of recovery, the current state of evidence regarding the outcome enhancing effects of CST Fb seem to suggest the possible combination of direct treatment effects *as well as* the indirect effect mediated by increased rate of patient retention in the treatment.

NOT Fb group and NOT P/T groups also showed statistically significant differences in the rate of recovery in some of the multilevel analyses. Considering the findings from the meta- and mega-analytic studies demonstrating the outcome enhancement associated with the NOT P/T Fb and NOT Fb groups, non-significant findings in terms of the rates of change may also support the notion of the retention effect that may mediate improvement in outcome. Future studies explicitly testing such a hypothesis may help further our understanding about the mechanisms of change. Although an explicit demonstration of this mediation effect was not tested in this study, combinations of aforementioned evidence seem to support this relationship.

To the extent that retaining patients in treatment plays an important role in enhancing treatment outcome among those who are predicted to experience negative outcome was supported, this notion presents some important clinical considerations for mental health care

systems and providers. The differences in the distribution of session attendance after the first signal alarm event across treatment groups and the mean differences in session attendance after the first signal alarm event suggest that, if patients are not identified as at risk of treatment failure and this information is not provided to therapists, patients seem more likely to leave treatment before experiencing improvement. Thus, implementation of routine monitoring of patient outcome, especially with a system that is capable of predicting negative outcome, appears important for at risk patients. Systems of care should seriously consider implementing such systems. Another clinical implication appears to be the importance of allowing patients to receive the care needed to achieve the desired improvement. The pooled dataset from the six major studies showed that the average number of session attendance for NOT patients was twice as many as that of all patients pooled together and that this had important positive consequences for patients.

Although natural log transformed time fitted the data better than the linear time, there was still a considerable amount of variation around the fitted trajectory both within and between patients. Such variation appears reflective of the complex patterns in which patients change through the course of therapy. In future studies, exploration of additional level-1 time varying predictors may be helpful in capturing the dynamic nature of patient change throughout the course of therapy.

Some limitations of this study should be noted. Although statistically significant findings favoring feedback treatment groups on a combined dataset were found, the same findings were not replicated in analyses where patients were grouped by the number of sessions attended after the first signal warning. Given the decreased sample size, however, separate analyses of the rate of change on the stratified samples of not-on-track patients may have been due to lack of

statistical power. Due to practical limitations associated with conducting numerous analyses, subscale score analyses were fitted to the statistical models that were fitted to the OQ total scale scores, which excluded some predictors that did not contribute to increased model fit. It is possible that building multilevel models of change "from scratch" for each of the subscales may find those excluded predictors to behave in a different manner. Another criticism may be made regarding the limited number of alternative models of change tested in this study. Although complex polynomial models of change, such as the quadratic and cubic models could not be tested on the entire duration of treatment when accounting for other change parameters (e.g., discontinuous intercept and slope) due to the technical limitations with the computer program used in this study, statistical modeling of the rate of change from the first signal warning event to termination could have included polynomial models of change. Thus, future examination of alternative models of change may lead to the development of models that better account for the change among the not-on-track patients.

Summary and Concluding Discussions

This meta-analytic and multi-level analytic review summarized the findings from the past six major feedback studies published to date. Exhaustive statistical analyses were conducted to obtain estimates of feedback effects in both the intent-to-treat and efficacy samples. These sets of analyses provided estimated effects when feedback interventions are implemented as a policy as well as when feedback interventions are evaluated among those who satisfied the least criteria to have likely benefitted from the feedback interventions. The results yielded generally smaller effects in the intent-to-treat analyses than in the efficacy analyses, however, both sets of analyses showed clinically significant treatment effects of the feedback interventions among those patients who were predicted to experience treatment failure. Such clinical benefits included

greater degree of distress reduction, decreased odds of deterioration, increased odds of achieving clinically significant improvement, and in some cases faster rates of recovery after first identified as at-risk cases. These benefits were found at the overall treatment outcome as measured by the OQ total scale scores as well as at more specific domains of the outcome as measured by the OQ subscale scores (i.e., symptom distress, interpersonal relations, and social role performance).

This quantitative review also found the unique contributions the newer forms of feedback intervention strategies (i.e., CST feedback and patient/therapist feedback) made to the patient outcome in relation to providing patient progress feedback to the therapists alone. This review also highlighted a retention effect in the newer feedback intervention, that is, *feedback interventions likely helped patients stay longer in treatment, which in turn contributed to improved outcome among at risk patients*. Limitations of this line of research were also discussed in the discussion sections of respective studies. Despite the limitations already discussed, however, the accumulating evidence appears substantial in favor of the routine use of progress feedback and clinical problem-solving tools. When considering clinicians' difficulty with identifying patients at risk of treatment failure (Hannan et al 2005), the current state of evidence seems sufficient to warrant routine use of these feedback interventions.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

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

American Psychological Association. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77*, 203-211.

Beckstead, D., J., Hatch, A. L., Lambert, M. J., Eggett, D. L., Goats, M. K., & Vermeersch, D. A. (2003). Clinical significance of the Outcome Questionnaire (OQ-45.2). *The Behavior Analyst Today, 4*, 79-90.

Beutler, L. E. (2001). Comparisons among quality assurance systems: From outcome assessment to clinical utility. *Journal of Consulting and Clinical Psychology, 69*, 197-204.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Cristoph, P., et al. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist, 51*, 3-16.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Cristoph, P., et al. (1996). An update on empirically validated therapies. *The Clinical Psychologist, 49*, 5-18.

DeRubeis, R. J., Gelfand, L. A., Tang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. *American Journal of Psychiatry, 156*, 1007-1013.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). An introduction to latent variable growth curve modeling: Concepts, issues, and applications (2nd Ed.). Lawrence Erlbaum Associates, Publishers: Mahwah.

Durham, C. J. (1999). Outcome questionnaire: Repeated administrations, mechanical responding, and social desirability (Doctoral dissertation, Brigham Young University, 1998). *Dissertation Abstracts International: Section B. Sciences and Engineering, 59(11)*, 6112.

Finch, A. 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 & Psychotherapy, 8*, 231-242.

Gallo, R. & Tasca, G. A. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: II. The complexities. *Psychotherapy Research, 19*, 438-452.

Grove, W. M. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of Clinical Psychology, 61*, 1233-1243.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., et al. (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., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9*, 329-343.

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

Hatfield, D., McCullough, L., Plucinski, A., & Krieger, K. (2009). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy*. Advance online publication. Doi:10.1002/cpp.656

*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.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

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.

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.

Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart-Heidelberg Model. *Journal of Consulting and Clinical Psychology, 69*, 173-183.

Lambert, M. J., Bergin, A. E., & Garfield, S. L. (2004). Introduction and historical overview. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.). New York: Wiley.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159-172.

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., Morton, J.J., Hatfield, D.R., Harmon, C., Hamilton, S., Reid, R.C. et al. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)*. American Professional Credentialing Services, L.L.C.

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 & 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.

*Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., et al. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: a replication. *Clinical Psychology & Psychotherapy, 9*, 91-103.

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.

Lunnen, C., & Ogles, B. M. (1998). A multi-perspective, multi-variable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 40-410.

Lutz, W., Lowry, J., Kopta, S. M., Einstein, D. A., & Howared, 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.

Mash, E. J., & Hunsley, J. (1993). Assessment considerations in the identification of failing psychotherapy: Bringing the negatives out of the darkroom. *Psychological Assessment, 5*, 292-301.

National Institutes of Health. (2002). State implementation of evidence-based practices: Bridging science and service (National Institute of Mental Health and Substance Abuse and Mental Health Services Administration Request for Application MH-03-007). Retrieved on December 19, 2006, from http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-03-007.html

Orwin, R, G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8,* 157-159.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354-379.

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

Reed, G. M., & Eisman, E. J. (2006). Uses and misuses of evidence: Managed care, treatment guidelines, and outcomes measurement in professional practice. In C. D. Goodheart, A. E. Kazdin, & R. J. Sternberg (Eds.), Evidence-based psychotherapy: Where practice and research meet (pp. 13-35). Washington, DC: American Psychological Association.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 85,* 638-664.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59-82.

Sapyta, J., Riemer, M, & Bickman, L. (2005). Feedback to clinicians: Theory, research, and practice. *Journal of Clinical Psychology, 62,* 145-153.

Serretti, A., Cusin, C., Rausch, J. J., Bondy, B., & Smeraldi, E. (2006). Pooling pharmacogenetic studies on the serotonin transporter: A mega-analysis. *Psychiatry Research, 145,* 61-65.

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.

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

*Slade, K., Lambert, M. J., Harmon, S. C., 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.

Spielmans, G. I., Masters, K. S., & Lambert, M. J. (2006). A comparison of rational versus empirical methods in the prediction of psychotherapy outcome. *Clinical Psychology and Psychotherapy, 13*, 202-214.

Umphress, V. J., Lambert, M. J., Smart, D. W., & Barlow, S. H. (1997). Concurrent and construct validity of the outcome questionnaire. *Journal of Psychoeducational Assessment, 15*, 40-55.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome questionnaire: Item sensitivity to cange. *Journal of Personality Assessment, 74*, 242-261.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology, 51*, 38-49.

Wendt, D. C., & Slife, B. D. (2007). Is evidence-based practice diverse enough? Philosophy of science considerations. *American Psychologist, 62*, 613–614.

*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 failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*, 59-68.

Appendix A

Tables and Figure

Table 1

Characteristics of Clients from Studies Used in Meta-Analyses

Study	Clients/therapists[a] N	Age M (SD)	Females %	Caucasians %	Dosage *M* (*SD*)	Intake OQ *M* (*SD*)	NOT[b] *n* (%)
Lambert, et al. (2001)	609/36	22.23 (3.92)	70.0	87.4	4.68 (3.89)	69.23 (23.20)	66 (10.8)
Lambert, et al. (2002)	1422/56	22.37 (3.74)	66.7	85.0	4.49 (3.39)	69.87 (22.58)	240 (16.9)
Whipple et al. (2003)	1339/49	23.01 (3.56)	63.5	86.0	5.14 (4.80)	69.27 (23.37)	278 (20.8)
Hawkins, et al. (2004)	306/5	30.51 (10.77)	63.1	94.1	6.06 (6.45)	83.23 (23.74)	101 (33.0)
Harmon, et al. (2007)	1374/72	22.68 (3.68)	61.0	83.0	6.74 (6.44)	71.23 (22.61)	369 (26.9)
Slade, et al. (2008)	1101/73	24.25 (3.29)	57.5	82.7	5.81 (5.67)	71.50 (22.07)	328 (29.8)

Note. [a]Numbers of clients and therapists prior to applying any exclusion criteria. Thus, the numbers reported here do not match with those reported in the original articles for studies that employed exclusion criteria i.e., Lambert et al. (2002), Whipple et al. (2003), and Hawkins, et al. (2004). [b]NOT = Clients whose progress was identified by OQ-45 algorithms as being Not-On-Track.

Table 2

Meta-Analysis of Effects of Feedback Interventions on Mean Post-Test OQ-45 Total Scale Score

Comparison	*k*	*n* grp1/grp2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent to Treat Analyses			
CST Fb vs. Fb	3	415 / 246	-0.16* [-0.33, -0.002]	0	0	-0.16 (0)
P/T Fb vs. Fb	3	222 / 188	-0.16 [-0.36, 0.03]	0	0	-0.16 (0)
Fb vs. TAU	4	269 / 318	-0.28** [-0.47, -0.10]	6	2	-0.28 (0)
CST Fb vs. TAU[a]	-	415 / 318	-0.44*** [-0.59, -0.30]	-	-	-
P/T Fb vs. TAU[b]	-	222 / 318	-0.36*** [-0.54, -0.19]	-	-	-
			Efficacy Analyses			
CST Fb vs. Fb	3	217 / 169	-0.19 [-0.43, 0.05]	0	1	-0.19 (0)
P/T Fb vs. Fb	3	177 / 147	-0.16 [-0.37, 0.06]	0	0	-0.11 (1)
Fb vs. TAU	4	136 / 318	-0.53*** [-0.78, -0.28]	20	8	-0.67 (2)
CST Fb vs. TAU[a]	-	217 / 318	-0.70*** [-0.88, -0.52]	-	-	-
P/T Fb vs. TAU[b]	-	177 / 318	-0.55*** [-0.73, -0.36]	-	-	-

Note. *k* = number of studies; *n* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level. Dashes in table indicate values are not applicable because given analysis was based on mega-analysis. [a]Mega-analysis using pooled CST Fb group versus pooled TAU group. [b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3

Clinical Significance Classification of Not-On-Track Patients by Treatment Conditions

	NOT				OT		
Clinical Significance	CST Fb	P/T Fb	Fb	TAU	P/T Fb	OT Fb	TAU
Treatment Conditions (Intent-To-Treat Sample)							
Worsened/Deteriorated	47 (11.3%)	35 (15.8%)	58 (13.6%)	64 (20.1%)	20 (2.1%)	45 (1.9%)	43 (3.0%)
No Change	212 (51.1%)	101 (45.5%)	237 (55.5%)	183 (57.5%)	507 (54.2%)	1485 (62.1%)	940 65.1%)
Improved/Recovered	156 (37.6%)	86 (38.7%)	132 (30.9%)	71 (22.3%)	408 (43.6%)	860 (36.0%)	461 (31.9%)
Treatment Conditions (Efficacy Sample)							
Worsened/Deteriorated	12 (5.5%)	26 (14.7%)	24 (9.1%)	-	20 (2.6%)	40 (2.4%)	-
No Change	91 (41.9%)	71 (40.1%)	140 (53.2%)	-	349 (44.9%)	794 (48.1%)	-
Improved/Recovered	114 (52.5%)	80 (45.2%)	99 (37.6%)	-	408 (52.5%)	817 (49.5%)	-

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. OT = patients whose progress was identified by OQ algorithms as being on track with expected recovery. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual.

Table 4

Meta-Analysis and Mega-Analysis of Effects of Feedback Interventions on Treatment Outcome:

Combined Odds Ratio of Reliable Worsening/Deterioration

Comparison	*k*	*OR* [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
		Intent to Treat Analyses			
CST Fb vs. Fb	3	0.76 [0.46, 1.26]	0	0	0.76 (0)
P/T Fb vs. Fb	3	1.35 [0.76, 2.41]	0	0	1.35 (0)
Fb vs. TAU	4	0.62* [0.40, 0.98]	3	1	0.70 (2)
CST Fb vs. TAU[a]	-	0.51** [0.34, 0.76]	-	-	-
P/T Fb vs. TAU[b]	-	0.74 [0.47, 1.17]	-	-	-
		Efficacy Analyses			
CST Fb vs. Fb	3	0.66 [0.29, 1.52]	0	1	0.83 (2)
P/T Fb vs. Fb	3	1.89 [0.90, 3.96]	0	1	2.95 (2)
Fb vs. TAU	4	0.44* [0.23, 0.85]	3	4	0.58 (2)
CST Fb vs. TAU[a]	-	0.23*** [0.12, 0.44]	-	-	-
P/T Fb vs. TAU[b]	-	0.68 [0.42, 1.13]	-	-	-

Note. *k* = number of studies; *OR* = Combined odds ratio (random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of null studies (with odds ratio of 1.00) needed to bring the combined odds ratio (fixed model) to above 0.66. Odds ratios smaller than 1.00 indicate lower odds of client deterioration among patients in the treatment group, favoring the treatment group in comparison to a control group. Dashes in table indicate values are not applicable because given analysis was based on mega-analysis.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 5

Meta-Analysis and Mega-Analysis of Effects of Feedback Interventions on Treatment Outcome:

Combined Odds Ratio of Clinically Significant Improvement

Comparison	*k*	*OR* [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
		Intent to Treat Analyses			
CST Fb vs. Fb	3	1.53* [1.08, 2.18]	2	1	1.40 (1)
P/T Fb vs. Fb	3	1.44 [0.95, 2.19]	0	0	1.44 (0)
Fb vs. TAU	4	1.70** [1.17, 2.46]	3	2	1.72 (1)
CST Fb vs. TAU[a]	-	2.01*** [1.51, 2.92]	-	-	-
P/T Fb vs. TAU[b]	-	2.20*** [1.51, 3.21]	-	-	-
		Efficacy Analyses			
CST Fb vs. Fb	3	1.83 [0.89, 3.76]	4	2	1.83 (0)
P/T Fb vs. Fb	3	1.38 [0.88, 2.18]	0	1	1.25 (2)
Fb vs. TAU	4	2.55*** [1.64, 3.98]	11	6	2.33 (1)
CST Fb vs. TAU[a]	-	3.85*** [2.65, 5.60]	-	-	-
P/T Fb vs. TAU[b]	-	2.97*** [1.93, 4.27]	-	-	-

Note. *k* = number of studies; *OR* = Combined odds ratio (random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of null studies (with odds ratio of 1.00) needed to bring the combined odds ratio (fixed model) to above 1.5. Odds ratios greater than 1.00 indicate higher odds of client improvement. Dashes in table indicate values are not applicable because given analysis was based on mega-analysis.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6

Meta-Analysis of Effects of Feedback Interventions on Pre-Post Change on OQ-45 Total Scale Score

Comparison	*n*	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
		Intent-to-treat analysis			
CST Fb	415	-0.45*** [-0.59, -0.31]	53	4	-0.45 (0)
NOT P/T Fb	222	-0.49*** [-0.75, -0.22]	28	4	-0.49 (0)
NOT Fb	427	-0.25*** [-0.33, -0.16]	45	2	-0.23 (1)
NOT TAU	318	-0.04 [-0.23, 0.15]	0	0	-0.04 (0)
		Efficacy analysis			
CST Fb	217	-0.82*** [-0.98, -0.66]	75	10	-0.82 (0)
NOT P/T Fb	177	-0.68*** [-1.04, -0.32]	34	7	-0.68 (0)
NOT Fb	263	-0.42*** [-0.58, -0.26]	81	7	-0.33 (2)

Note. k = number of studies; n = number of participants; ES = weighted effect size (Hedges's g; random effect model); CI = Confidence Interval; Classic fail-safe N = the number of null studies needed to bring the combined p-value to above 0.05 (two-tail); Orwin's fail-safe N = the number of studies (with null mean Hedges's g) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level. *p < .05. **p < .01. ***p < .001.

Table 7

Meta-Analysis of Effects of Feedback Interventions on Mean Difference in OQ-45 Total Scale Pre-Post Change Scores

Comparison	*k*	*n* grp1/grp2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent-to-treat analyses			
CST Fb vs. Fb	3	415 / 246	-0.21* [-0.38, -0.04]	3	1	-0.21 (0)
P/T Fb vs. Fb	3	222 / 188	-0.17 [-0.37, 0.03]	0	0	-0.17 (0)
Fb vs. TAU	4	269 / 318	-0.28*** [-0.44, -0.11]	6	2	-0.26 (1)
CST Fb vs. TAU[a]	-	415 / 318	-0.43*** [-0.58, -0.32]	-	-	-
P/T Fb vs. TAU[b]	-	222 / 318	-0.44*** [-0.62, -0.27]	-	-	-
			Efficacy analyses			
CST Fb vs. Fb	3	217 / 169	-0.29 [-0.60, 0.02]	4	2	-0.19 (0)
P/T Fb vs. Fb	3	177 / 147	-0.15 [-0.37, 0.07]	0	0	-0.11 (1)
Fb vs. TAU	4	136 / 318	-0.60*** [-0.81, -0.40]	27	9	-0.65 (1)
CST Fb vs. TAU[a]	-	217 / 318	-0.78*** [-0.96, -0.60]	-	-	-
P/T Fb vs. TAU[b]	-	177 / 318	-0.56*** [-0.74, -0.37]	-	-	-

Note. *k* = number of studies; *n* = number of participants; ES = weighted effect size (Hedges's *g*; random effect model) ; CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 8

Meta-Analysis and Mega-Analysis of Effects of Feedback Interventions on Mean Number of Session Attendance

Comparison	*k*	*n* gr1/gr2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent-to-treat analyses			
CST Fb vs. Fb	3	415/ 246	0.41* [0.05, 0.76]	14	3	0.41 (0)
P/T Fb vs. Fb	3	222/ 188	0.12 [-0.11, 0.35]	0	0	0.12 (0)
Fb vs. TAU	4	269/ 318	0.27 [-0.16, 0.70]	4	0	0.42 (1)
CST Fb vs. TAU[a]	-	415/ 318	0.48*** [0.33, 0.63]	-	-	-
P/T Fb vs. TAU[b]	-	222/ 318	0.40*** [0.23, 0.58]	-	-	-

Note. *k* = number of studies; *n* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model) ; gr = group; CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail; based on fixed model); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to below 0.2. Positive effect sizes indicate more number of session attendances among patients in the treatment group in comparison to the control group. Dashes in table indicate values are not applicable because given analysis was based on mega-analysis.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 9

Comparison of Mean OQ Total Scores at Pre-Treatment, at the Time of Signal Warning, and Post-Treatment by Treatment Conditions

Treatment Condition		Original Study Dataset			Merged Dataset		
		Pre-Test	Warning	Post-Test	Pre-Test	Warning	Post-Test
CST Fb	*M*	79.8	88.2	70.5	79.4	87.8	70.8
	(*SD*)	(19.3)	(15.0)	(21.5)	(19.8)	(15.2)	(21.6)
	n	415	415	415	412	409	412
NOT P/T Fb	*M*	81.4	89.4	72.0	81.0	88.8	72.6
	(*SD*)	(19.0)	(15.2)	(22.9)	(18.9)	(16.0)	(22.9)
	n	222	222	222	221	218	221
NOT Fb	*M*	79.8	88.9	74.2	79.4	88.9	74.1
	(*SD*)	(20.5)	(16.4)	(23.6)	(20.4)	(16.1)	(23.1)
	n	427	427	427	423	418	423
NOT TAU	*M*	80.3	90.3	80.0	80.4	90.3	79.7
	(*SD*)	(20.2)	(15.8)	(21.0)	(20.1)	(15.8)	(21.1)
	n	318	318	318	312	309	312
OT P/T Fb	*M*	69.1		54.8	68.8		54.6
	(*SD*)	(23.1)		(23.4)	(23.1)		(23.4)
	n	934		934	915		915
OT Fb	*M*	67.7		56.1	67.6		55.9
	(*SD*)	(23.3)		(22.9)	(23.2)		(22.7)
	n	2387		2387	2346		2346
OT TAU	*M*	68.2		58.9	67.9		58.5
	(*SD*)	(22.8)		(22.6)	(22.7)		(22.6)
	n	1442		1442	1414		1414
Total	*M*	70.8	89.1	60.6	70.6	88.8	60.4
	(*SD*)	(23.0)	(15.7)	(24.0)	(22.9)	(15.8)	(24.0)
	N	6145	1382	6145	6044	1354	6044

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. OT = patients whose progress was identified by OQ algorithms as being on track with expected recovery. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual. The *n* for pre-treatment and post-treatment are matched because of the use of the last observation carried forward method, where missing post-treatment scores were replaced by the last scores recorded.

Table 10

Correlations of Pre-treatment, Signal-warning, Post-treatment, and Pre-post Change Scores between the OQ Total Scale and OQ Subscales

OQ Total Scale	SD scale	IR scale	SR scale
Pre-treatment score	.95	.75	.77
Score at signal warning*	.90	.60	.63
Post-treatment score	.96	.81	.81
Pre-post Change score	.94	.76	.75

Note. *Correlations for scores at signal warning were based on not-on-track patients only (n = 1068). All other correlations were based on all patients with merged subscale scores (N = 6044). All correlations were significant at $p < 0.001$.

Table 11

Meta-Analysis of Effects of Feedback Interventions on Mean Post-Test OQ-45 Symptom Distress (SD) Subscale Score

Feedback Condition	*k*	*N* grp1/grp2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent to Treat Analyses			
CST Fb vs. Fb	3	412 / 244	-0.15 [-0.32, 0.01]	0	0	-0.15 (0)
P/T Fb vs. Fb	3	221 / 186	-0.11 [-0.30, 0.09]	0	0	-0.08 (1)
Fb vs. TAU	4	266 / 312	-0.26** [-0.42, -0.09]	5	2	-0.26 (0)
CST Fb vs. TAU[a]	-	412 / 312	-0.33*** [-0.47, -0.18]	-	-	-
P/T Fb vs. TAU[b]	-	221 / 312	-0.24** [-0.41, -0.06]	-	-	-
			Efficacy Analyses			
CST Fb vs. Fb	3	214 / 168	-0.18 [-0.38, 0.03]	0	0	-0.18 (0)
P/T Fb vs. Fb	3	176 / 147	-0.09 [-0.31, 0.13]	0	0	-0.09 (0)
Fb vs. TAU	4	135 / 312	-0.52*** [-0.73, -0.32]	18	7	-0.60 (2)
CST Fb vs. TAU[a]	-	214 / 312	-0.55*** [-0.73, -0.38]	-	-	-
P/T Fb vs. TAU[b]	-	176 / 312	-0.39*** [-0.57, -0.20]	-	-	-

Note. *k* = number of studies; *N* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 12

Meta-Analysis of Effects of Feedback Interventions on Mean Post-Test OQ-45 Interpersonal Relations (IR) Subscale Score

Feedback Condition	*k*	*N* grp1/grp2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent to Treat Analyses			
CST Fb vs. Fb	3	412 / 244	-0.08 [-0.24, 0.09]	0	0	-0.15 (0)
P/T Fb vs. Fb	3	221 / 186	-0.06 [-0.28, 0.15]	0	0	-0.06 (0)
Fb vs. TAU	4	266 / 312	-0.24* [-0.43, -0.05]	3	0	-0.37 (2)
CST Fb vs. TAU[a]	-	412 / 312	-0.42*** [-0.57, -0.27]	-	-	-
P/T Fb vs. TAU[b]	-	221 / 312	-0.29** [-0.46, -0.12]	-	-	-
			Efficacy Analyses			
CST Fb vs. Fb	3	214 / 168	-0.12 [-0.32, 0.08]	0	0	-0.12 (0)
P/T Fb vs. Fb	3	176 / 147	-0.12 [-0.34, 0.10]	0	0	-0.12 (0)
Fb vs. TAU	4	135 / 312	-0.37* [-0.67, -0.07]	9	5	-0.53 (2)
CST Fb vs. TAU[a]	-	214 / 312	-0.54*** [-0.71, -0.36]	-	-	-
P/T Fb vs. TAU[b]	-	176 / 312	-0.43*** [-0.61, -0.24]	-	-	-

Note. *k* = number of studies; *N* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 13

Meta-Analysis of Effects of Feedback Interventions on Mean Post-Test OQ-45 Social Role (SR) Subscale Score

Feedback Condition	*k*	*N* grp1/grp2	ES [95% CI]	Classic failsafe *N*	Orwin's failsafe *N*	Trim and Fill ES (studies trimmed)
			Intent to Treat Analyses			
CST Fb vs. Fb	3	412 / 244	-0.20* [-0.36, -0.03]	2	0	-0.08 (2)
P/T Fb vs. Fb	3	221 / 186	-0.22 [-0.52, 0.07]	1	1	-0.22 (0)
Fb vs. TAU	4	266 / 312	-0.24** [-0.40, -0.07]	6	2	-0.35 (2)
CST Fb vs. TAU[a]	-	412 / 312	-0.42*** [-0.57, -0.27]	-	-	-
P/T Fb vs. TAU[b]	-	221 / 312	-0.40*** [-0.57, -0.23]	-	-	-
			Efficacy Analyses			
CST Fb vs. Fb	3	214 / 168	-0.28** [-0.49, -0.08]	3	2	-0.36 (2)
P/T Fb vs. Fb	3	176 / 147	-0.24 [-0.61, 0.12]	1	1	-0.24 (0)
Fb vs. TAU	4	135 / 312	-0.48*** [-0.68, -0.21]	12	6	-0.58 (2)
CST Fb vs. TAU[a]	-	214 / 312	-0.54*** [-0.71, -0.36]	-	-	-
P/T Fb vs. TAU[b]	-	176 / 312	-0.56*** [-0.75, -0.37]	-	-	-

Note. *k* = number of studies; *N* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model); CI = Confidence Interval; Classic fail-safe *N* = the number of null studies needed to bring the combined *p*-value to above 0.05 (two-tail); Orwin's fail-safe *N* = the number of studies (with null mean Hedges's *g*) needed to bring the combined effect size (fixed model) to above -0.2. Negative effect sizes indicate lower distress level.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 14

Summary of Effects of Feedback Interventions on Mean Post-Test OQ-45 Subscale Scores

Feedback Condition	*k*	*N* grp1/grp2	SD scale ES [95% CI]	IR scale ES [95% CI]	SR scale ES [95% CI]
			Intent-to-treat analyses		
CST Fb vs. Fb	3	412 / 244	-0.15 [-0.32, 0.01]	-0.08 [-0.24, 0.09]	-0.20* [-0.36, -0.03]
P/T Fb vs. Fb	3	221 / 186	-0.11 [-0.30, 0.09]	-0.06 [-0.28, 0.15]	-0.22 [-0.52, 0.07]
Fb vs. TAU	4	266 / 312	-0.26** [-0.42, -0.09]	-0.24* [-0.43, -0.05]	-0.24** [-0.40, -0.07]
CST Fb vs. TAU	-	412 / 312	-0.33*** [-0.47, -0.18]	-0.42*** [-0.57, -0.27]	-0.42*** [-0.57, -0.27]
P/T Fb vs. TAU	-	221 / 312	-0.24** [-0.41, -0.06]	-0.29** [-0.46, -0.12]	-0.40*** [-0.57, -0.23]
			Efficacy Analyses		
CST Fb vs. Fb	3	214 / 168	-0.18 [-0.38, 0.03]	-0.12 [-0.32, 0.08]	-0.28** [-0.49, -0.08]
P/T Fb vs. Fb	3	176 / 147	-0.09 [-0.31, 0.13]	-0.12 [-0.34, 0.10]	-0.24 [-0.61, 0.12]
Fb vs. TAU	4	135 / 312	-0.52*** [-0.73, -0.32]	-0.37* [-0.67, -0.07]	-0.48*** [-0.68, -0.21]
CST Fb vs. TAU[a]	-	214 / 312	-0.55*** [-0.73, -0.38]	-0.54*** [-0.71, -0.36]	-0.54*** [-0.71, -0.36]
P/T Fb vs. TAU[b]	-	176 / 312	-0.39*** [-0.57, -0.20]	-0.43*** [-0.61, -0.24]	-0.56*** [-0.75, -0.37]

Note. *k* = number of studies; *N* = total number of participants; ES = weighted effect size (Hedges's *g*; random effect model); CI = Confidence Interval; Negative effect sizes indicate lower distress level.
[a]Mega-analysis using pooled CST Fb group versus pooled TAU group.
[b]Mega-analysis using pooled P/T Fb group versus pooled TAU group.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 15

Mega-Analysis of Effects of Feedback Interventions on Pre-Post Change on OQ-45 Subscale Scores

Comparison	*N*	SD scale ES [95% CI]	IR scale ES [95% CI]	SR scale ES [95% CI]
		Intent-to-treat analyses		
CST Fb	412	-0.43*** [-0.53, -0.33]	-0.23*** [-0.33, -0.13]	-0.29*** [-0.40, -0.17]
NOT P/T Fb	221	-0.42*** [-0.56, -0.27]	-0.16* [-0.30, -0.02]	-0.37*** [-0.51, -0.23]
NOT Fb	423	-0.30*** [-0.39, -0.21]	-0.06 [-0.15, 0.03]	-0.13** [-0.23, -0.03]
NOT TAU	312	-0.14* [-0.25, -0.02]	0.12* [0.02, 0.22]	0.09 [-0.03, 0.21]
		Efficacy analysis		
CST Fb	214	-0.76*** [-0.91, -0.60]	-0.46*** [-0.61, -0.32]	-0.54*** [-0.70, -0.38]
NOT P/T Fb	176	-0.56*** [-0.74, -0.37]	-0.28** [-0.44, -0.11]	-0.50*** [-0.68, -0.32]
NOT Fb	262	-0.49*** [-0.61, -0.37]	-0.17** [-0.29, -0.06]	-0.25*** [-0.39, -0.12]
NOT TAU	312	-	-	-

Note. k = number of studies; N = total number of participants; ES = pre-post change effect size (Hedges's g); CI = Confidence Interval; Negative effect sizes indicate lower distress level.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 16

Mega-Analysis of Effects of Feedback Interventions on Mean Difference in OQ Subscale Pre-Post Change Scores

Comparison	*N* grp1/grp2	SD scale ES [95% CI]	IR scale ES [95% CI]	SR scale ES [95% CI]
		Intent-to-treat analyses		
CST Fb vs. Fb	412 / 244	-0.13 [-0.29, 0.03]	-0.12 [-0.28, 0.04]	-0.18* [-0.34, -0.02]
P/T Fb vs. Fb	221 / 186	-0.10 [-0.29, 0.10]	-0.07 [-0.26, 0.13]	-0.29** [-0.48, -0.09]
Fb vs. TAU	266 / 312	-0.20* [-0.37, -0.04]	-0.18* [-0.35, -0.02]	-0.28*** [-0.44, -0.11]
CST Fb vs. TAU	412 / 312	-0.30*** [-0.45, -0.15]	-0.36*** [-0.51, -0.21]	-0.34*** [-0.49, -0.19]
P/T Fb vs. TAU	221 / 312	-0.27** [-0.44, -0.10]	-0.29*** [-0.47, -0.12]	-0.44*** [-0.62, -0.27]
		Efficacy Analyses		
CST Fb vs. Fb	214 / 168	-0.23* [-0.44, -0.03]	-0.29** [-0.49, -0.09]	-0.27** [-0.47, -0.07]
P/T Fb vs. Fb	176 / 147	-0.08 [-0.29, 0.14]	-0.13 [-0.35, 0.09]	-0.24* [-0.46, -0.03]
Fb vs. TAU	135 / 312	-0.47*** [-0.68, -0.27]	-0.36*** [-0.56, -0.15]	-0.41*** [-0.61, -0.20]
CST Fb vs. TAU	214 / 312	-0.60*** [-0.78, -0.42]	-0.60*** [-0.78, -0.42]	-0.57*** [-0.74, -0.39]
P/T Fb vs. TAU	176 / 312	-0.39*** [-0.57, -0.20]	-0.41*** [-0.60, -0.22]	-0.53*** [-0.71, -0.34]

Note. k = number of studies; N = total number of participants; ES = weighted effect size (Hedges's g; random effect model); CI = Confidence Interval; Negative effect sizes indicate lower distress level.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 17

OQ Symptom Distress (SD) Subscale Score-Based Clinical Significance Classification of Patients by Treatment Conditions

	NOT				OT		
Clinical Significance	CST Fb	P/T Fb	Fb	TAU	P/T Fb	OT Fb	TAU
			Intent-To-Treat Sample				
Worsened/Deteriorated	45 (10.9%)	30 (13.6%)	51 (12.1%)	49 (15.7%)	22 (2.4%)	41 (1.7%)	33 (2.3%)
No Change	215 (52.2%)	109 (49.3%)	242 (57.2%)	172 (55.1%)	521 (56.9%)	1449 (61.7%)	936 (66.2%)
Improved/Recovered	152 (36.9%)	82 (37.1%)	130 (30.7%)	91 (29.2%)	372 (40.7%)	857 (36.5%)	445 (31.5%)
			Efficacy Sample				
Worsened/Deteriorated	13 (6.1%)	21 (11.9%)	22 (8.4%)	-	22 (2.8%)	33 (2.0%)	-
No Change	92 (43.0%)	80 (45.5%)	139 (53.1%)	-	379 (49.0%)	817 (49.7%)	-
Improved/Recovered	109 (50.9%)	75 (42.6%)	101 (38.5%)	-	372 (48.1%)	794 (48.3%)	-

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. OT = patients whose progress was identified by OQ algorithms as being on track with expected recovery. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual. The *n* for pre-treatment and post-treatment are matched because of the use of the last observation carried forward method, where missing post-treatment scores were replaced by the last scores recorded.

Table 18

OQ Interpersonal Relations (IR) Subscale Score-Based Clinical Significance Classification of Patients by Treatment Conditions

	NOT				OT		
Clinical Significance	CST Fb	P/T Fb	Fb	TAU	P/T Fb	OT Fb	TAU
			Intent-To-Treat Sample				
Worsened/Deteriorated	18 (4.4%)	15 (6.8%)	36 (8.5%)	30 (9.6%)	4 (0.4%)	30 (1.3%)	26 (1.8%)
No Change	334 (81.1%)	175 (79.2%)	347 (82.0%)	262 (84.0%)	764 (83.6%)	2035 (86.8%)	1249 (88.3%)
Improved/Recovered	60 (14.6%)	31 (14.0%)	40 (9.5%)	20 (6.4%)	146 (16.0%)	280 (11.9%)	139 (9.8%)
			(Efficacy Sample				
Worsened/Deteriorated	7 (3.3%)	11 (6.3%)	18 (6.9%)	-	4 (0.5%)	26 (1.6%)	-
No Change	163 (76.2%)	134 (76.1%)	211 (80.5%)	-	623 (80.6%)	1351 (82.2%)	-
Improved/Recovered	44 (20.6%)	31 (17.6%)	33 (12.6%)	-	146 (18.9%)	267 (16.3%)	-

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. OT = patients whose progress was identified by OQ algorithms as being on track with expected recovery. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual. The *n* for pre-treatment and post-treatment are matched because of the use of the last observation carried forward method, where missing post-treatment scores were replaced by the last scores recorded.

Table 19

OQ Social Role (SR) Subscale Score-Based Clinical Significance Classification of Patients by Treatment Conditions

	NOT				OT		
Clinical Significance	CST Fb	P/T Fb	Fb	TAU	P/T Fb	OT Fb	TAU
			Intent-to-treat sample				
Worsened/Deteriorated	18 (4.4%)	15 (6.8%)	36 (8.5%)	30 (9.6%)	10 (1.1%)	26 (1.1%)	25 (1.8%)
No Change	334 (81.1%)	175 (79.2%)	347 (82.0%)	262 (84.0%)	774 (84.6%)	2062 (87.9%)	1270 (89.8%)
Improved/Recovered	60 (14.6%)	31 (14.0%)	40 (9.5%)	20 (6.4%)	131 (14.3%)	259 (11.0%)	119 (8.4%)
			Efficacy sample				
Worsened/Deteriorated	5 (2.3%)	9 (5.1%)	14 (5.3%)	-	10 (1.3%)	25 (1.5%)	-
No Change	166 (77.6%)	135 (76.7%)	213 (81.3%)	-	632 (81.8%)	1375 (83.6%)	-
Improved/Recovered	43 (20.1%)	32 (18.2%)	35 (13.4%)	-	131 (16.9%)	244 (14.8%)	-

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. OT = patients whose progress was identified by OQ algorithms as being on track with expected recovery. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual. The *n* for pre-treatment and post-treatment are matched because of the use of the last observation carried forward method, where missing post-treatment scores were replaced by the last scores recorded.

Table 20

Mega-Analysis of Effects of Feedback Interventions on Reducing Deterioration at Termination in Not-on-Track (NOT) Patients Based on OQ Subscale Scores

Comparison	SD scale *OR* [95% CI]	IR scale *OR* [95% CI]	SR scale *OR* [95% CI]
Intent-to-treat analyses			
CST Fb vs. Fb	0.87 [0.53, 1.43]	0.65 [0.33, 1.30]	0.51 [0.25, 1.01]
P/T Fb vs. Fb	1.11 [0.62, 1.99]	0.77 [0.37, 1.61]	0.64 [0.28, 1.45]
Fb vs. TAU	0.66 [0.40, 1.07]	0.93 [0.53, 1.64]	0.64 [0.32, 1.25]
CST Fb vs. TAU	0.66 [0.43, 1.02]	0.43** [0.23, 0.79]	0.46* [0.24, 0.88]
P/T Fb vs. TAU	0.84 [0.52, 1.38]	0.68 [0.36, 1.30]	0.62 [0.28, 1.36]
Efficacy analyses			
CST Fb vs. Fb	0.77 [0.35, 1.71]	0.60 [0.22, 1.64]	0.42 [0.14, 1.29]
P/T Fb vs. Fb	1.40 [0.67, 2.90]	0.82 [0.35, 1.96]	0.94 [0.35, 2.49]
Fb vs. TAU	0.43* [0.21, 0.88]	0.67 [0.31, 1.46]	0.53 [0.21, 1.33]
CST Fb vs. TAU	0.35** [0.18, 0.66]	0.32** [0.14, 0.74]	0.27** [0.10, 0.73]
P/T Fb vs. TAU	0.73 [0.42, 1.26]	0.63 [0.31, 1.28]	0.62 [0.28, 1.36]

Note. OR = odds ratio; CI = Confidence Interval; Odds ratios smaller than 1.00 indicate lower odds of client deterioration among patients in the treatment group, favoring the treatment group in comparison to the control group. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 21

Mega-Analysis of Effects of Feedback Interventions on Enhancing Clinically Significant Improvement in Not-on-Track (NOT) Patients at Termination based on OQ Subscale Scores

Comparison	SD scale *OR* [95% CI]	IR scale *OR* [95% CI]	SR scale *OR* [95% CI]
Intent-to-treat analyses			
CST Fb vs. Fb	1.32 [0.94, 1.85]	1.43 [0.88, 2.33]	1.67* [1.00, 2.78]
P/T Fb vs. Fb	1.34 [0.88, 2.02]	1.43 [0.78, 2.63]	1.87 [0.99, 3.51]
Fb vs. TAU	1.10 [0.77, 1.57]	1.45 [0.78, 2.68]	2.16** [1.25, 3.72]
CST Fb vs. TAU	1.42* [1.04, 1.95]	2.49*** [1.47, 4.23]	2.18** [1.32, 3.62]
P/T Fb vs. TAU	1.43 [0.99, 2.07]	2.38** [1.32, 4.30]	2.21** [1.26, 3.87]
Efficacy analyses			
CST Fb vs. Fb	1.69* [1.12, 2.54]	1.81* [1.03, 3.19]	1.97* [1.10, 3.53]
P/T Fb vs. Fb	1.28 [0.82, 2.00]	1.53 [0.82, 2.87]	1.96* [1.01, 3.77]
Fb vs. TAU	1.72* [0.13, 2.62]	2.25* [1.15, 4.40]	2.86*** [1.56, 5.24]
CST Fb vs. TAU	2.52*** [1.75, 3.62]	3.78*** [2.16, 6.62]	3.16*** [1.84, 5.42]
P/T Fb vs. TAU	1.80** [1.23, 2.65]	3.12*** [1.72, 5.67]	2.79*** 1.58, 4.95]

Note. OR = odds ratio; CI = Confidence Interval; Odds ratios greater than 1.00 indicate higher odds of client improvement among patients in the treatment group, favoring the treatment group in comparison to the control group. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 22

Number of Sessions Attended by Not-On-Track Patients after the First Signal Warning Event

# of sessions after first worsening	Treatment group among NOT patients			
	CST Fb	P/T Feedback	Feedback	No feedback (TAU)
1 – 2 sessions	92 (22.2%)	52 (23.4%)	135 (31.6%)	111 (34.9%)
3 – 4 sessions	84 (20.2%)	41 (18.5%)	93 (21.8%)	71 (22.3%)
5 – 8 sessions	101 (24.3%)	55 (24.8%)	101 (23.7%)	84 (26.4%)
9 – 11 sessions	53 (12.8%)	24 (10.8%)	42 (9.8%)	27 (8.5%)
12 – 18 sessions	51 (12.3%)	33 (14.9%)	43 (10.1%)	20 (6.3%)
> 19 sessions	34 (8.2%)	17 (7.7%)	13 (3.0%)	5 (1.6%)

Note. NOT = patients whose progress was identified by OQ algorithms as being not on track. CST Fb = NOT patients whose therapists received clinical support tools feedback in addition to OQ progress feedback. P/T Fb = both patients and therapists received patient OQ progress feedback. Fb = patients whose therapists received patient OQ progress feedback. TAU = treatment as usual.

Table 23

Comparison of Linear Models of Change on ITT Sample

Parameter		Model A[a]	Model B.1[b]	Model C.1[c]	Model D.1[d]	Model E.1[e]	Model F.1[f]
				Fixed effects			
Initial Status	Intercept	78.25*** (0.45)	81.04*** (0.51)	80.29*** (0.53)	79.06*** (0.52)	78.67*** (0.53)	78.70*** (0.52)
	#Sessions[g]			-2.84*** (0.79)			
	Signal[h]					10.72*** (0.38)	11.77*** (0.39)
Rate of change	Session		-0.62*** (0.06)	-0.55*** (0.07)	0.68*** (0.09)	-1.41*** (0.09)	-1.42*** (0.09)
	SessionPW[i]				-2.17*** (0.14)	-0.97*** (0.13)	-0.17*** (0.14)
	Session ×#Sessions			-0.19 (0.12)			
	SessionPW ×#SessionsPW[j]						2.36*** (0.17)
				Random effects			
Level 1	Within-person	177.38*** (2.32)	139.25*** (1.94)	139.05*** (1.93)	132.52*** (1.88)	115.53*** (1.70)	115.48*** (1.69)
Level 2	Initial status	256.69*** (10.78)	303.56*** (22.48)	298.48*** (13.32)	307.23*** (14.12)	318.93*** (14.30)	318.89*** (14.30)
	Session		2.72*** (0.22)	2.78*** (0.23)	1.38*** (0.28)	1.33*** (0.22)	1.31*** 0.21)
	SesssionPW				4.18*** (0.73)	5.33*** (0.76)	3.68*** (0.63)
	Signal					53.85*** (6.88)	55.68*** (6.90)
-2 log likelihood		108580.6	107032.0	107007.2	106658.7	105604.1	105430.6
AIC		108586.6	107044.0	107023.2	106678.7	105634.1	105462.6
BIC		108609.0	107088.8	107083.0	106753.5	105746.3	105582.3

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on sample mean; h = intercept at the time of first signal warning, noting the elevation of intercept in OQ points; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track patients centered on sample mean
** $p < .01$, *** $p < .001$.

Table 24

Comparison of Alternative Linear Models of Change (Efficacy sample)

Parameter		Model A[a]	Model B.1[b]	Model C.1[c]	Model D.1[d]	Model E.1[e]	Model F.1[f]
				Fixed effects			
Initial Status	Intercept	77.58*** (0.52)	82.53*** (0.66)	82.19*** (0.66)	80.92*** (0.58)	80.92*** (0.58)	78.27*** (0.59)
	#Sessions[g]			-0.57 (1.11)			
	Signal[h]					10.13*** (0.45)	11.22*** (0.46)
Rate of change	Session		-0.79 *** (0.07)	-0.85*** (0.08)	0.52*** (0.11)	-1.50*** (0.10)	-1.50*** (0.10)
	SessionPW[i]				-2.03*** (0.15)	-0.75*** (0.14)	-1.68*** (0.16)
	Session ×#Sessions			0.29* (0.14)			
	SessionPW ×#SessionsPW[j]						2.41*** (0.18)
				Random effects			
Level 1	Within-person	183.63*** (2.70)	141.99*** (2.21)	142.20*** (2.22)	135.36*** (2.15)	120.57*** (1.98)	120.22*** (1.96)
Level 2	Initial status	244.21*** (12.18)	282.95*** (15.03)	278.78*** (14.85)	296.17*** (16.32)	305.76*** (16.48)	305.92*** (16.50)
	Session		2.54*** (0.23)	2.48*** (0.22)	1.64*** (0.35)	1.42*** (0.25)	1.37*** (0.24)
	SesssionPW				3.54*** (0.74)	4.29*** (0.72)	3.11*** (0.61)
	Signal					50.52*** (7.85)	52.06*** (7.87)
-2 log likelihood		84875.7	83517.4	83508.6	83254.0	82574.1	82416.4
AIC		84881.7	83529.4	83524.6	83274.0	82604.1	82448.4
BIC		84903.4	83572.8	83582.4	83346.4	82712.6	82564.1

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on sample mean; h = intercept at the time of first signal warning, noting the elevation of intercept in OQ points; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track patients centered on sample mean
** $p < .01$, *** $p < .001$.

Table 25

Comparison of Alternative Linear Models of Change (ITT Sample without Last Observation Carried Forward Method)

Parameter		Model A[a]	Model B.1[b]	Model C.1[c]	Model D.1[d]	Model E.1[e]	Model F.1[f]
				Fixed effects			
Initial Status	Intercept	77.96*** (0.45)	81.05*** (0.50)	80.57*** (0.52)	79.18*** (0.52)	78.69*** (0.52)	78.69*** (0.52)
	#Sessions[g]			-2.99*** (0.79)			
	Signal[h]					10.37*** (0.38)	11.40*** (0.38)
Rate of change	Session		-0.73 *** (0.06)	-0.74*** (0.07)	0.49*** (0.09)	-1.43*** (0.09)	-1.42*** (0.09)
	SessionPW[i]				-2.05*** (0.14)	-0.99 *** (0.13)	-1.96*** (0.14)
	Session ×#Sessions			0.19 (0.12)			
	SessionPW ×#SessionsPW[j]						2.83*** (0.18)
				Random effects			
Level 1	Within-person	179.05*** (2.37)	140.89*** (1.98)	141.06*** (1.98)	134.01*** (1.92)	118.42*** (1.76)	117.68*** (1.74)
Level 2	Initial status	253.94*** (10.77)	298.99*** (13.41)	293.74*** (16.39)	304.60*** (14.11)	316.80*** (14.33)	316.78*** (14.33)
	Session		2.55*** (0.21)	2.51*** (0.21)	1.16*** (0.25)	1.32*** (0.22)	1.28*** (0.22)
	SesssionPW				4.10*** (0.71)	5.26*** (0.77)	3.75*** (0.64)
	Signal					44.29*** (6.62)	44.82*** (6.56)
-2 log likelihood		106201.7	104672.4	104658.2	104320.2	103392.5	103158.9
AIC		106207.7	104684.4	104674.2	104340.2	103422.5	103190.9
BIC		106230.1	104729.2	104733.9	104414.8	103534.3	103310.2

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on sample mean; h = intercept at the time of first signal warning, noting the elevation of intercept in OQ points; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track patients centered on sample mean
** $p < .01$, *** $p < .001$.

Table 26

Comparison of Alternative Linear Models of Change (Efficacy Sample without Last Observation Carried Forward Method)

Parameter		Model A[a]	Model B.1[b]	Model C.1[c]	Model D.1[d]	Model E.1[e]	Model F.1[f]
				Fixed effects			
Initial Status	Intercept	77.42*** (0.52)	81.30*** (0.58)	81.47*** (0.58)	79.29** (0.61)	78.86*** (0.61)	78.85*** (0.61)
	#Sessions[g]			-2.78** (0.79)			
	Signal[h]					9.93*** (0.44)	11.02*** (0.45)
Rate of change	Session		-0.84 *** (0.07)	-0.97*** (0.08)	0.42** (0.11)	-1.51*** (0.10)	-1.50*** (0.11)
	SessionPW[i]				-1.96** (0.15)	-0.76 *** (0.14)	-1.78*** (0.16)
	Session ×#Sessions			0.50*** (0.14)			
	SessionPW ×#SessionsPW[j]						2.59*** (0.19)
				Random effects			
Level 1	Within-person	184.12*** (2.72)	142.63*** (2.23)	142.89*** (2.23)	135.96*** (2.17)	121.45*** (2.00)	121.20*** (1.99)
Level 2	Initial status	242.11*** (12.14)	279.18*** (14.91)	275.11*** (14.73)	292.71*** (16.22)	303.76*** (16.47)	303.84*** (16.46)
	Session		2.43*** (0.22)	2.32*** (0.21)	1.45*** (0.33)	1.40*** (0.25)	1.33*** (0.24)
	SesssionPW				3.50*** (0.73)	4.28*** (0.72)	3.17*** (0.61)
	Signal					44.95*** (7.63)	45.44** (7.60)
-2 log likelihood		84031.0	82684.1	82669.1	82428.3	81793.6	81615.5
AIC		84040.0	82696.1	82685.1	82448.3	81823.6	81647.5
BIC		84061.7	82739.4	82742.9	82520.6	81931.9	81763.1

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on sample mean; h = intercept at the time of first signal warning, noting the elevation of intercept in OQ points; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track patients centered on sample mean
** $p < .01$, *** $p < .001$.

Table 27

Comparison of Alternative Linear Models of Change Based on Natural Log Transformed Time (ITT Sample)

Parameter		Model A[a]	Model B.2[b]	Model C.2[c]	Model D.2[d]	Model E.2[e]	Model F.2[f]
				Fixed effects			
Initial Status	Intercept	78.25*** (0.45)	81.79*** (0.56)	81.09*** (0.59)	80.74*** (0.50)	77.56*** (0.50)	77.61*** (0.50)
	#Sessions[g]			-0.82 (0.89)			
	Signal[h]					11.03*** (0.34)	11.42*** (0.35)
Rate of change	Session				0.09 (0.07)	-.20*** (0.04)	-0.24*** (0.05)
	LgSession		-2.27*** (0.26)	-2.05*** (0.26)			
	LgSessionPW[i]				-3.97*** (0.41)	-7.96*** (0.37)	-8.19*** (0.36)
	LgSession ×#Sessions			-2.62*** (0.42)			
	LgSessionPW ×#SessionsPW[j]						2.42*** (0.45)
				Random effects			
Level 1	Within-person	177.38*** (2.32)	146.70*** (2.02)	146.18*** (2.01)	130.88*** (1.85)	117.99*** (1.71)	118.12*** (1.71)
Level 2	Initial status	256.69*** (10.78)	348.31*** (16.72)	347.67*** (16.69)	305.11*** (13.38)	293.89*** (13.00)	293.33*** (12.98)
	Signal					32.15*** (5.82)	33.26*** (5.87)
	Session				1.10*** (0.18)		
	LgSession		47.25*** (3.26)	47.08*** (3.21)			
	LgSesssionPW				77.93*** (7.71)	63.18*** (4.75)	57.47*** (4.53)
-2 log likelihood		108580.6	107460.1	107388.3	106567.3	105419.8	105393.0
AIC		108586.6	107472.1	107404.3	106587.3	105441.8	105518.8
BIC		108609.0	107516.9	107464.1	106662.1	105524.1	105506.8

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model, using linear time throughout the course of treatment and natural log of time after the first signal warning; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on NOT sample mean; h = elevation in intercept at the time of first signal warning; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track (NOT) patients centered on NOT sample mean. ** $p < .01$, *** $p < .001$.

Table 28

Comparison of Alternative Linear Models of Change Based on Natural Log Transformed Time (Efficacy Sample)

Parameter		Model A[a]	Model B.2[b]	Model C.2[c]	Model D.2[d]	Model E.2[e]	Model F.2[f]
				Fixed effects			
Initial Status	Intercept	77.58*** (0.52)	82.53*** (0.66)	82.19*** (0.66)	80.92*** (0.58)	77.66*** (0.58)	77.74*** (0.58)
	#Sessions[g]			-0.57 (1.11)			
	Signal[h]					10.73*** (0.41)	11.23*** (0.42)
Rate of change	Session				0.06 (0.08)	-0.19*** (0.05)	-0.25*** (0.05)
	LgSession		-3.15 *** (0.29)	-2.85*** (0.30)			
	LgSessionPW[i]				-4.47*** (0.47)	-8.10*** (0.41)	-8.72*** (0.41)
	Session ×#Sessions			-1.70** (0.53)			
	SessionPW ×#SessionsPW[j]						3.50*** (0.54)
				Random effects			
Level 1	Within-person	183.63*** (2.70)	149.63*** (2.32)	149.42*** (2.31)	133.24*** (2.11)	122.33*** (1.98)	122.38*** (1.98)
Level 2	Initial status	244.21*** (12.18)	329.24*** (18.88)	328.82*** (18.87)	285.19*** (14.96)	279.20*** (14.85)	278.46*** (14.83)
	Signal					32.51*** (6.93)	34.18*** (7.00)
	Session				1.11*** (0.21)		
	LgSession		47.62*** (3.72)	47.85*** (3.72)			
	LgSesssionPW				76.39*** (8.50)	57.38*** (4.90)	50.84*** (1.54)
-2 log likelihood		84875.7	83842.9	83823.5	83139.3	82376.7	82336.5
AIC		84881.7	83854.9	83839.5	83159.3	82398.7	82360.5
BIC		84903.4	83898.3	83897.3	83231.7	82478.2	82447.3

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model, using linear time throughout the course of treatment and natural log of time after the first signal warning; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on NOT sample mean; h = elevation in intercept at the time of first signal warning; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track (NOT) patients centered on NOT sample mean. ** $p < .01$, *** $p < .001$.

Table 29

Comparison of Alternative Linear Models of Change Based on Natural Log Transformed Time (ITT Sample without Last Observation Carried Forward Method)

Parameter		Model A[a]	Model B.2[b]	Model C.2[c]	Model D.2[d]	Model E.2[e]	Model F.2[f]
				Fixed effects			
Initial Status	Intercept	77.96*** (0.45)	82.09*** (0.57)	81.44*** (0.59)	80.57*** (0.50)	77.52*** (0.50)	77.61*** (0.50)
	#Sessions[g]			-1.17 (0.90)			
	Signal[h]					10.71*** (0.34)	11.32*** (0.35)
Rate of change	Session				0.12 (0.07)	-0.18*** (0.04)	-0.26*** (0.05)
	LgSession		-2.79*** (0.26)	-2.53*** (0.26)			
	LgSessionPW[i]				-4.86*** (0.43)	-8.43*** (0.38)	-9.18*** (0.37)
	Session ×#Sessions			-1.74*** (0.44)			
	SessionPW ×#SessionsPW[j]						5.20*** (0.51)
				Random effects			
Level 1	Within-person	179.05*** (2.37)	148.46*** (2.07)	148.07*** (2.06)	131.40*** (1.87)	120.15*** (1.76)	119.86*** (1.75)
Level 2	Initial status	253.94*** (10.77)	364.57*** (16.89)	345.81*** (16.87)	301.79*** (13.33)	292.09*** (13.00)	291.05*** (12.96)
	Signal					24.84*** (5.63)	26.07*** (5.67)
	Session				2.84*** (0.59)		
	LgSession		46.04*** (3.29)	46.64*** (3.30)			
	LgSesssionPW				82.76*** (8.19)	66.50*** (5.10)	56.88*** (4.55)
-2 log likelihood		106201.7	105082.2	105046.4	104167.2	103147.1	103047.6
AIC		106207.7	105094.2	105062.4	104187.2	103169.1	103071.6
BIC		106230.1	105138.9	105122.1	104261.8	103251.2	103161.1

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model, using linear time throughout the course of treatment and natural log of time after the first signal warning; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on NOT sample mean; h = elevation in intercept at the time of first signal warning; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track (NOT) patients centered on NOT sample mean. ** $p < .01$, *** $p < .001$.

Table 30

Comparison of Alternative Linear Models of Change Based on Natural Log Transformed Time (Efficacy Sample without Last Observation Carried Forward Method)

Parameter		Model A[a]	Model B.2[b]	Model C.2[c]	Model D.2[d]	Model E.2[e]	Model F.2[f]
				Fixed effects			
Initial Status	Intercept	77.43*** (0.52)	82.68** (0.66)	82.41*** (0.66)	80.84*** (0.58)	77.62*** (0.58)	77.73*** (0.58)
	#Sessions[g]			-0.85 (1.11)			
	Signal[h]					10.56*** (0.41)	11.16*** (0.41)
Rate of change	Session				0.07 (0.08)	-0.18*** (0.05)	-0.26*** (0.05)
	Intercept		-3.41*** (0.30)	-3.17*** (0.31)			
	SessionPW[i]				-4.83*** (0.47)	-8.30*** (0.41)	-9.26*** (0.41)
	Session ×#Sessions			-1.08*** (0.54)			
	SessionPW ×#SessionsPW[j]						4.88*** (0.56)
				Random effects			
Level 1	Within-person	184.12*** (2.37)	150.38*** (2.33)	150.23*** (2.33)	133.36*** (2.12)	123.28*** (2.00)	123.10*** (1.99)
Level 2	Initial status	242.11*** (12.14)	326.37*** (18.88)	325.75*** (18.86)	281.75*** (14.83)	277.57*** (14.81)	276.61*** (14.77)
	Signal					27.88*** (6.75)	28.92*** (6.79)
	Session				1.08*** (0.21)		
	LgSession		46.09*** (3.67)	46.64*** (3.68)			
	LgSesssionPW				78.60*** (8.68)	58.66*** (5.04)	50.97*** (4.56)
-2 log likelihood		84034.0	83002.0	82992.2	82289.2	81572.8	81500.9
AIC		84040.0	83014.0	83008.2	82309.2	81594.8	81524.9
BIC		84061.7	83057.4	83066.0	82381.5	51674.3	81611.6

Note. Standard errors are in parenthesis. a = unconditional means model; b = Unconditional linear growth model; c = Stratified linear growth model; d = Discontinuous slope model, using linear time throughout the course of treatment and natural log of time after the first signal warning; e = Discontinuous intercept and slope model; f = Stratified discontinuous intercept and slope model; g = natural log transformation of number of sessions attended by not-on-track patients centered on NOT sample mean; h = elevation in intercept at the time of first signal warning; i = rate of change after the first warning signal event; j = natural log transformation of number of sessions attended after first signal warning by not-on-track (NOT) patients centered on NOT sample mean. ** $p < .01$, *** $p < .001$.

Table 31

Comparison of Discontinuous Intercept and Slope Models (Symptom Distress Subscale)

Parameter	Model F.1[a]	Model F.2[b]
	Fixed effects	
Initial Status		
Intercept	46.83*** (0.35)	46.32*** (0.34)
Signal[c]	6.93*** (0.26)	7.02*** (0.25)
Rate of chang[e]		
Session[d]	-1.03*** (0.07)	-0.41*** (0.05)
SessionPW[e]	-1.04*** (0.10)	
LgSessionPW[f]		-4.97*** (0.27)
SessionPW × #SessionsPW	1.71*** (0.11)	3.56*** (0.33)
	Random effects	
Level 1		
Within-person	49.06*** (0.74)	46.95*** (0.72)
Level 2		
Initial status	142.61*** (6.47)	138.37*** (6.16)
Signal	21.70*** (3.06)	21.17*** (3.06)
Session	0.67*** (0.12)	0.76*** (0.12)
SessionPW	1.53*** (0.29)	
LgSesssionPW		27.56*** (3.26)
-2 log likelihood	90424.7	90006.8
AIC	90456.7	90038.8
BIC	90575.7	90157.8

Note: Standard errors are in parenthesis. a = stratified discontinuous intercept and slope model, using linear time from intake to termination and another linear time from the first warning signal event to termination; b = stratified discontinuous intercept and slope model, using linear time from intake to termination and an additional slope from the time of first signal warning to termination; a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event; c = time variant indicating if a given observation occurred before or after the first signal warning; d = session number; session number after the first warning; natural log transformation of session numbers after the first signal warning event. ** $p < .01$, *** $p < .001$.

Table 32
Comparison of Discontinuous Intercept and Slope Models (Interpersonal Relations Subscale)

Parameter	Model F.1[a]	Model F.2[b]
	Fixed effects	
Initial Status		
Intercept	16.75*** (0.15)	16.63*** (0.15)
Signal[c]	2.74*** (0.12)	2.87*** (0.12)
Rate of chang[e]		
Session[d]	-0.26*** (0.03)	-0.12*** (0.03)
SessionPW[e]	-0.40*** (0.05)	
LgSessionPW[f]		-1.62*** (0.12)
SessionPW × #SessionsPW	0.54*** (0.05)	1.23*** (0.14)
	Random effects	
Level 1		
Within-person	10.12*** (0.15)	9.66*** (0.15)
Level 2		
Initial status	26.93*** (1.25)	26.12*** (1.18)
Signal	3.59*** (0.67)	5.04*** (0.66)
Session	0.16*** (0.04)	0.15*** (0.03)
SessionPW	0.37*** (0.07)	
LgSesssionPW		6.28*** (0.71)
-2 log likelihood	70389.4	70117.6
AIC	70421.4	70149.6
BIC	70540.4	70268.7

Note: Standard errors are in parenthesis. a = stratified discontinuous intercept and slope model, using linear time from intake to termination and another linear time from the first warning signal event to termination; b = stratified discontinuous intercept and slope model, using linear time from intake to termination and an additional slope from the time of first signal warning to termination; a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event; c = time variant indicating if a given observation occurred before or after the first signal warning; d = session number; session number after the first warning; natural log transformation of session numbers after the first signal warning event. ** $p < .01$, *** $p < .001$.

Table 33

Comparison of Discontinuous Intercept and Slope Models (Social role performance Subscale)

Parameter	Model F.1[a]	Model F.2[b]
	Fixed effects	
Initial Status		
Intercept	14.94*** (0.11)	14.87*** (0.11)
Signal[c]	1.94*** (0.10)	2.20*** (0.09)
Rate of chang[e]		
Session[d]	-0.17*** (0.03)	-0.08*** (0.02)
SessionPW[e]	-0.39*** (0.04)	
LgSessionPW[f]		-1.52*** (0.10)
SessionPW × #SessionsPW	0.41*** (0.03)	0.76*** (0.11)
	Random effects	
Level 1		
Within-person	7.29*** (0.11)	7.05*** (0.11)
Level 2		
Initial status	12.64*** (0.63)	12.32*** (0.60)
Signal	2.21*** (0.40)	2.22*** (0.40)
Session	0.07*** (0.02)	0.08*** (0.01)
SessionPW	0.24*** (0.04)	
LgSesssionPW		3.30*** (0.45)
-2 log likelihood	65392.7	65145.2
AIC	65424.7	65177.2
BIC	65543.7	65296.2

Note: Standard errors are in parenthesis. a = stratified discontinuous intercept and slope model, using linear time from intake to termination and another linear time from the first warning signal event to termination; b = stratified discontinuous intercept and slope model, using linear time from intake to termination and an additional slope from the time of first signal warning to termination; a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event; c = time variant indicating if a given observation occurred before or after the first signal warning; d = session number; session number after the first warning; natural log transformation of session numbers after the first signal warning event. ** $p < .01$, *** $p < .001$.

Table 34

Multilevel Analyses of Rate of Change in OQ Total Scale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (ITT Analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	79.43** (0.46)	84.62** (0.44)	86.67** (0.96)
CST Fb			-1.64 (1.23)
P/T Fb			-2.30 (1.42)
Fb			-1.06 (1.22)
#SessionsPW-1[a]			-1.90** (0.62)
Rate of change[b]			
Intercept (TAU)		-2.26** (0.10)	-2.90** (0.23)
SessionPW (CST Fb)			-0.67* (0.27)
SessionPW (P/T Fb)			-0.52 (0.31)
SessionPW (Fb)			-0.55* (0.28)
#SessionsPW-1			2.86** (0.18)
		Random effects	
Level 1			
Within-person	191.50** (3.20)	135.17** (2.44)	133.78** (2.40)
Level 2			
Initial status	237.17** (11.05)	203.70** (10.25)	201.56** (10.06)
Rate of change after warning		4.92** (0.50)	3.33** (0.36)
-2 log likelihood	71345.9	69682.7	69423.6
AIC	71351.9	69694.7	69451.6
BIC	71373.0	69737.0	69550.3

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 35

Multilevel Analyses of Rate of Change in OQ Total Scale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (ITT Analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	79.43** (0.46)	87.79** (0.45)	89.03** (0.98)
CST Fb			-1.23 (1.27)
P/T Fb			-1.25 (1.47)
Fb			-0.61 (1.25)
#SessionsPW-1[a]			-0.64 (.63)
Rate of change (log)[b]			
Intercept (TAU)		-9.00** (0.33)	-8.61** (0.74)
SessionPW (CST Fb)			-1.95* (0.92)
SessionPW (P/T Fb)			-2.07 (1.05)
SessionPW (Fb)			-1.91* (0.94)
#SessionsPW-1			5.19** (0.54)
		Random effects	
Level 1			
Within-person	191.50** (3.20)	132.06** (2.37)	131.85** (2.36)
Level 2			
Initial status	237.17** (11.05)	189.85** (10.71)	189.52** (10.67)
Rate of change after warning		58.55** (4.85)	50.88** (4.42)
-2 log likelihood	71345.9	69333.9	69239.7
AIC	71351.9	69345.9	69267.7
BIC	71373.0	69388.2	69366.4

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 36

Multilevel Analyses of Rate of Change in OQ Total Scale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (Efficacy Analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	78.19** (0.54)	84.06** (0.51)	86.99** (0.98)
CST Fb			-1.80 (1.51)
P/T Fb			-3.64* (1.53)
Fb			2.18 (1.38)
#SessionsPW-1[a]			-1.26 (0.83)
Rate of change[b]			
Intercept (TAU)		-2.14** (0.10)	-2.82** (0.22)
SessionPW (CST Fb)			-0.79** (0.27)
SessionPW (P/T Fb)			-0.44 (0.29)
SessionPW (Fb)			-0.49 (0.28)
#SessionsPW-1			2.66** (0.19)
		Random effects	
Level 1			
Within-person	192.48** (3.48)	136.84** (2.67)	135.77** (2.63)
Level 2			
Initial status	236.39** (12.67)	197.06** (11.61)	195.56** (11.41)
Rate of change after warning		4.23** (0.47)	2.83** (0.34)
-2 log likelihood	59275.0	57875.1	57672.0
AIC	59281.0	57887.1	57700.0
BIC	59301.5	57928.3	57796.1

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 37

Multilevel Analyses of Rate of Change in OQ Total Scale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformed Time (Efficacy Analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	78.19** (0.54)	87.80** (0.53)	89.28** (1.00)
CST Fb			-0.39 (1.55)
P/T Fb			-2.32 (1.59)
Fb			-1.42 (1.42)
#SessionsPW-1[a]			-0.22 (0.85)
Rate of change (log)[b]			
Intercept (TAU)		-8.90** (0.34)	-8.56** (0.72)
SessionPW (CST Fb)			-3.00** (0.98)
SessionPW (P/T Fb)			-1.83 (1.04)
SessionPW (Fb)			-1.74 (0.97)
#SessionsPW-1			4.83** (0.61)
		Random effects	
Level 1			
Within-person	192.48** (3.48)	133.59** (2.58)	133.59** (2.57)
Level 2			
Initial status	236.39** (12.67)	180.80** (12.15)	179.85** (12.08)
Rate of change after warning		51.78** (4.79)	45.01 (4.38)
-2 log likelihood	59275.0	57565.8	57495.1
AIC	59281.0	57577.8	57523.1
BIC	59301.5	57619.0	57619.2

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 38

Results of Analysis of Rate of Change on OQ Symptom Distress (SD) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (ITT analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	46.71** (0.31)	49.82** (0.31)	50.72** (0.67)
CST Fb			-0.47 (0.85)
P/T Fb			-0.88 (0.99)
Fb			-0.05 (0.85)
#SessionsPW-1[a]			-0.78 (0.43
Rate of change[b]			
Intercept (TAU)		-1.31** (0.07)	-1.74** (0.15)
SessionPW (CST Fb)			-0.41* (0.17)
SessionPW (P/T Fb)			-0.29 (0.19)
SessionPW (Fb)			-0.38* (0.18)
#SessionsPW-1			1.70** (0.11)
		Random effects	
Level 1			
Within-person	78.00** (1.32)	55.73** (1.01)	55.25** (1.00)
Level 2			
Initial status	109.20** (5.03)	100.04** (4.88)	99.96** (4.84)
Rate of change after warning		1.84** (0.19)	1.26** (0.14)
-2 log likelihood	62408.4	60882.8	60659.0
AIC	62414.4	60894.8	60687.0
BIC	62435.5	60937.0	60785.3

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 39

Results of Analysis of Rate of Change on OQ Symptom Distress (SD) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformed Time (ITT analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	46.71** (0.31)	51.80** (0.32)	52.24** (0.69)
CST Fb			-0.27 (0.86)
P/T Fb			-0.23 (1.02)
Fb			0.23 (0.88)
#SessionsPW-1[a]			0.06 (0.88)
Rate of change (log)[b]			
Intercept (TAU)		-5.51** (0.21)	-5.32** (0.48)
SessionPW (CST Fb)			-1.10 (0.60)
SessionPW (P/T Fb)			-1.13 (0.68)
SessionPW (Fb)			-1.26* (0.61)
#SessionsPW-1			3.04** (0.35)
		Variance components	
Level 1			
Within-person	78.00** (1.32)	54.22** (0.99)	54.14** (0.98)
Level 2			
Initial status	109.20** (5.03)	97.63** (5.20)	97.90** (5.20)
Rate of change after warning		23.81** (1.99)	21.11** (1.84)
-2 log likelihood	62408.4	60575.1	50494.6
AIC	62414.4	60587.1	60522.6
BIC	62435.5	60629.3	60620.9

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 40

Results of Analysis of Rate of Change on OQ Symptom Distress (SD) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (Efficacy analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	45.88** (0.37)	49.46** (0.36)	50.95** (0.68)
CST Fb			-0.76 (1.06)
P/T Fb			-1.78 (1.07)
Fb			-0.77 (0.96)
#SessionsPW-1[a]			-0.34 (0.58)
Rate of change[b]			
Intercept (TAU)		-1.31** (0.07)	-1.71*** (0.14)
SessionPW (CST Fb)			-0.49** (0.17)
SessionPW (P/T Fb)			-0.24 (0.19)
SessionPW (Fb)			-0.36* (0.18)
#SessionsPW-1			1.60** (0.12)
		Random effects	
Level 1			
Within-person	78.50** (1.43)	56.16** (2.44)	55.81** (1.20)
Level 2			
Initial status	109.43** (5.80)	98.36** (5.60)	98.44** (5.56)
Rate of change after warning		1.64** (0.19)	1.11** (0.14)
-2 log likelihood	51721.4	50418.8	50241.6
AIC	51727.4	50430.8	50269.6
BIC	51747.9	50471.9	50365.3

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 41

Results of Analysis of Rate of Change on OQ Symptom Distress (SD) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (Efficacy analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	45.88** (0.37)	51.81** (0.37)	52.42** (0.71)
CST Fb			0.24 (1.10)
P/T Fb			-0.97 (1.12)
Fb			-0.28 (1.00)
#SessionsPW-1[a]			-0.40 (0.60)
Rate of change (log)[b]			
Intercept (TAU)		-5.52** (0.23)	-5.30** (0.48)
SessionPW (CST Fb)			-1.78** (0.64)
SessionPW (P/T Fb)			-0.98 (0.68)
SessionPW (Fb)			-1.21 (0.63)
#SessionsPW-1			2.85** (0.40)
		Variance components	
Level 1			
Within-person	78.50** (1.43)	54.49** (1.07)	54.49** (1.07)
Level 2			
Initial status	109.43** (5.80)	95.96** (6.01)	95.86** (6.00)
Rate of change after warning		21.96** (2.02)	19.51** (1.88)
-2 log likelihood	51721.4	50137.4	50076.8
AIC	51727.4	50149.4	50104.8
BIC	51747.9	50190.4	50200.5

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 42

Results of Analysis of Rate of Change on OQ Interpersonal Relations (IR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (ITT analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	17.53** (0.14)	18.47** (0.14)	19.17** (0.31)
CST Fb			-0.67 (0.40)
P/T Fb			-0.62 (0.46)
Fb			-0.57 (0.40)
#SessionsPW-1[a]			-0.28 (0.20)
Rate of change[b]			
Intercept (TAU)		-0.41** (0.03)	-0.52** (0.06)
SessionPW (CST Fb)			-0.22* (0.07)
SessionPW (P/T Fb)			-0.16* (0.08)
SessionPW (Fb)			-0.14 (0.08)
#SessionsPW-1			0.59** (0.05)
		Random effects	
Level 1			
Within-person	14.35** (0.24)	11.23** (0.20)	11.16** (0.20)
Level 2			
Initial status	22.40** (1.02)	22.17** (1.07)	20.04** (1.05)
Rate of change after warning		0.28** (0.03)	0.22** (0.03)
-2 log likelihood	48472.1	47528.0	47372.5
AIC	48478.1	47540.0	47400.5
BIC	48499.2	47582.2	47498.8

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 43

Results of Analysis of Rate of Change on OQ Interpersonal Relations (IR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (ITT analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	46.71** (0.31)	51.80** (0.32)	52.24** (0.69)
CST Fb			-0.27 (0.86)
P/T Fb			-0.23 (1.02)
Fb			0.23 (0.88)
#SessionsPW-1[a]			0.06 (0.88)
Rate of change (log)[b]			
Intercept (TAU)		-5.51** (0.21)	-5.32** (0.48)
SessionPW (CST Fb)			-1.10 (0.60)
SessionPW (P/T Fb)			-1.13 (0.68)
SessionPW (Fb)			-1.26* (0.61)
#SessionsPW-1			3.04** (0.35)
		Variance components	
Level 1			
Within-person	78.00** (1.32)	54.22** (0.99)	54.14** (0.98)
Level 2			
Initial status	109.20** (5.03)	97.63** (5.20)	97.90** (5.20)
Rate of change after warning		23.81** (1.99)	21.11** (1.84)
-2 log likelihood	62408.4	60575.1	50494.6
AIC	62414.4	60587.1	60522.6
BIC	62435.5	60629.3	60620.9

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 44

Results of Analysis of Rate of Change on OQ Interpersonal Relations (IR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (Efficacy Analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	17.37** (0.17)	18.47** (0.17)	19.27** (0.32)
CST Fb			-0.58 (0.50)
P/T Fb			-1.06* (0.51)
Fb			-0.71 (0.46)
#SessionsPW-1[a]			-0.18 (0.28)
Rate of change[b]			
Intercept (TAU)		-0.39** (0.03)	-0.50** (0.06)
SessionPW (CST Fb)			-0.25** (0.08)
SessionPW (P/T Fb)			-0.14 (0.08)
SessionPW (Fb)			-0.14 (0.08)
#SessionsPW-1			0.53** (0.05)
		Random effects	
Level 1			
Within-person	14.63** (0.27)	11.50** (0.23)	11.48** (0.23)
Level 2			
Initial status	23.52** (1.23)	22.14** (1.29)	22.99** (1.27)
Rate of change after warning		0.25** (0.03)	0.20** (0.03)
-2 log likelihood	40253.9	39464.8	39348.0
AIC	40259.9	39476.8	39376.0
BIC	40280.4	39517.8	39471.8

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 45

Results of Analysis of Rate of Change on OQ Interpersonal Relations (IR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (Efficacy Analysis).

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	17.37** (0.17)	19.18** (0.18)	19.64** (0.34)
CST Fb			-0.13 (0.53)
P/T Fb			-0.84 (0.54)
Fb			-0.52 (0.49)
#SessionsPW-1[a]			-0.52 (0.49)
Rate of change (log)[b]			
Intercept (TAU)		-1.69** (0.10)	-1.53** (0.21)
SessionPW (CST Fb)			-0.94** (0.29)
SessionPW (P/T Fb)			-0.43 (0.30)
SessionPW (Fb)			-0.43 (0.28)
#SessionsPW-1			1.08** (0.18)
		Random effects	
Level 1			
Within-person	14.63** (0.27)	11.27** (0.22)	11.28** (0.22)
Level 2			
Initial status	23.52** (1.23)	23.82** (1.43)	23.68** (1.42)
Rate of change after warning		4.09** (0.39)	3.72** (0.37)
-2 log likelihood	40253.9	39299.7	39248.6
AIC	40259.9	39311.7	39276.6
BIC	40280.4	39352.7	39372.4

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 46

Results of Analysis of Linear Rate of Change on OQ Social role performance (SR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (ITT analysis)

Parameter	Model G	Model H.1	Model I.1
	Fixed effects		
Initial Status			
Intercept (TAU)	15.21** (0.10)	16.13** (0.10)	16.68** (0.21)
CST Fb			-0.56* (0.27)
P/T Fb			-0.68* (0.32)
Fb			-0.53 (0.27)
#SessionsPW-1[a]			-0.70** (0.14)
Rate of change[b]			
Intercept (TAU)		-0.37** (0.02)	-0.52** (0.05)
SessionPW (CST Fb)			-0.08 (0.06)
SessionPW (P/T Fb)			-0.08 (0.08)
SessionPW (Fb)			-0.06 (0.08)
#SessionsPW-1			0.43** (0.04)
	Random effects		
Level 1			
Within-person	9.82** (0.17)	7.84** (0.14)	7.80** (0.14)
Level 2			
Initial status	9.84** (0.48)	9.57** (0.51)	9.08** (0.50)
Rate of change after warning		0.14** (0.02)	0.11** (0.01)
-2 log likelihood	44842.0	43864.4	43693.2
AIC	44848.0	43876.4	43721.2
BIC	44869.1	43918.6	43819.5

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 47

Results of Analysis of Log Linear Rate of Change on OQ Social role performance (SR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (ITT analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	15.21** (0.10)	16.74** (0.10)	17.10** (0.23)
CST Fb			-0.50 (0.29)
P/T Fb			-0.40 (0.34)
Fb			-0.44 (0.29)
#SessionsPW-1[a]			-0.46** (0.14)
Rate of change (log)[b]			
Intercept (TAU)		-1.58** (0.07)	-1.55** (0.16)
SessionPW (CST Fb)			-0.17 (0.20)
SessionPW (P/T Fb)			-0.46* (0.23)
SessionPW (Fb)			-0.26 (0.21)
#SessionsPW-1			0.78** (0.12)
		Random effects	
Level 1			
Within-person	9.82** (0.17)	7.69** (0.14)	7.69** (0.14)
Level 2			
Initial status	9.84** (0.48)	9.46** (0.60)	9.20** (0.60)
Rate of change after warning		2.17** (0.21)	1.97** (0.20)
-2 log likelihood	44842.0	43638.1	43580.5
AIC	44848.0	43650.1	43608.5
BIC	44869.1	43692.2	43706.8

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

Table 48

Results of Analysis of Rate of Change on OQ Social role performance (SR) Subscale Scores after First Signal Warning by Feedback Treatment Conditions Based on Linear Time (Efficacy Analysis)

Parameter	Model G	Model H.1	Model I.1
		Fixed effects	
Initial Status			
Intercept (TAU)	14.96** (0.11)	15.99** (0.11)	16.71** (0.22)
CST Fb			-0.71* (0.33)
P/T Fb			-0.76* (0.34)
Fb			-0.62* (0.31)
#SessionsPW-1[a]			-0.70* (0.14)
Rate of change[b]			
Intercept (TAU)		-0.34** (0.02)	-0.51** (0.05)
SessionPW (CST Fb)			-0.06 (0.06)
SessionPW (P/T Fb)			-0.07 (0.06)
SessionPW (Fb)			-0.04 (0.06)
#SessionsPW-1			0.41** (0.04)
		Random effects	
Level 1			
Within-person	9.84** (0.18)	7.94** (0.16)	7.90** (0.15)
Level 2			
Initial status	9.66** (0.54)	9.37** (0.59)	8.85** (0.56)
Rate of change after warning		0.12** (0.02)	0.09** (0.01)
-2 log likelihood	37120.9	36324.2	36189.8
AIC	37126.9	36336.2	36217.8
BIC	37127.4	36377.2	36313.6

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of session number after the first signal warning event.
* < .05, ** < .01

Table 49

Results of Analysis of Rate of Change on OQ Total Scale Scores after First Signal Warning by Feedback Treatment Conditions Based on Log Transformation of Time (Efficacy Analysis)

Parameter	Model G	Model H.2	Model I.2
		Fixed effects	
Initial Status			
Intercept (TAU)	14.96** (0.11)	16.7** (0.12)	17.12** (0.23)
CST Fb			-0.52 (0.36)
P/T Fb			-0.41 (0.37)
Fb			-0.48 (0.37)
#SessionsPW-1[a]			-0.46** (0.33)
Rate of change (log)[b]			
Intercept (TAU)		-1.56** (0.07)	-1.54** (0.16)
SessionPW (CST Fb)			-0.29 (0.21)
SessionPW (P/T Fb)			-0.44 (0.23)
SessionPW (Fb)			-0.22 (0.21)
#SessionsPW-1			0.72** (0.13)
		Random effects	
Level 1			
Within-person	9.84** (0.18)	7.78** (0.15)	7.78** (0.15)
Level 2			
Initial status	9.66** (0.54)	9.33** (0.66)	9.05** (0.65)
Rate of change after warning		1.93** (0.22)	1.74** (0.21)
-2 log likelihood	37120.9	36135.3	36092.4
AIC	37126.9	36147.3	36120.4
BIC	37127.4	36188.3	36216.2

Note: a = natural log transformation of the number of post-warning sessions (session-1) centered on NOT patient mean; b = rate of change as a function of natural log of session number after the first signal warning event.
* < .05, ** < .01

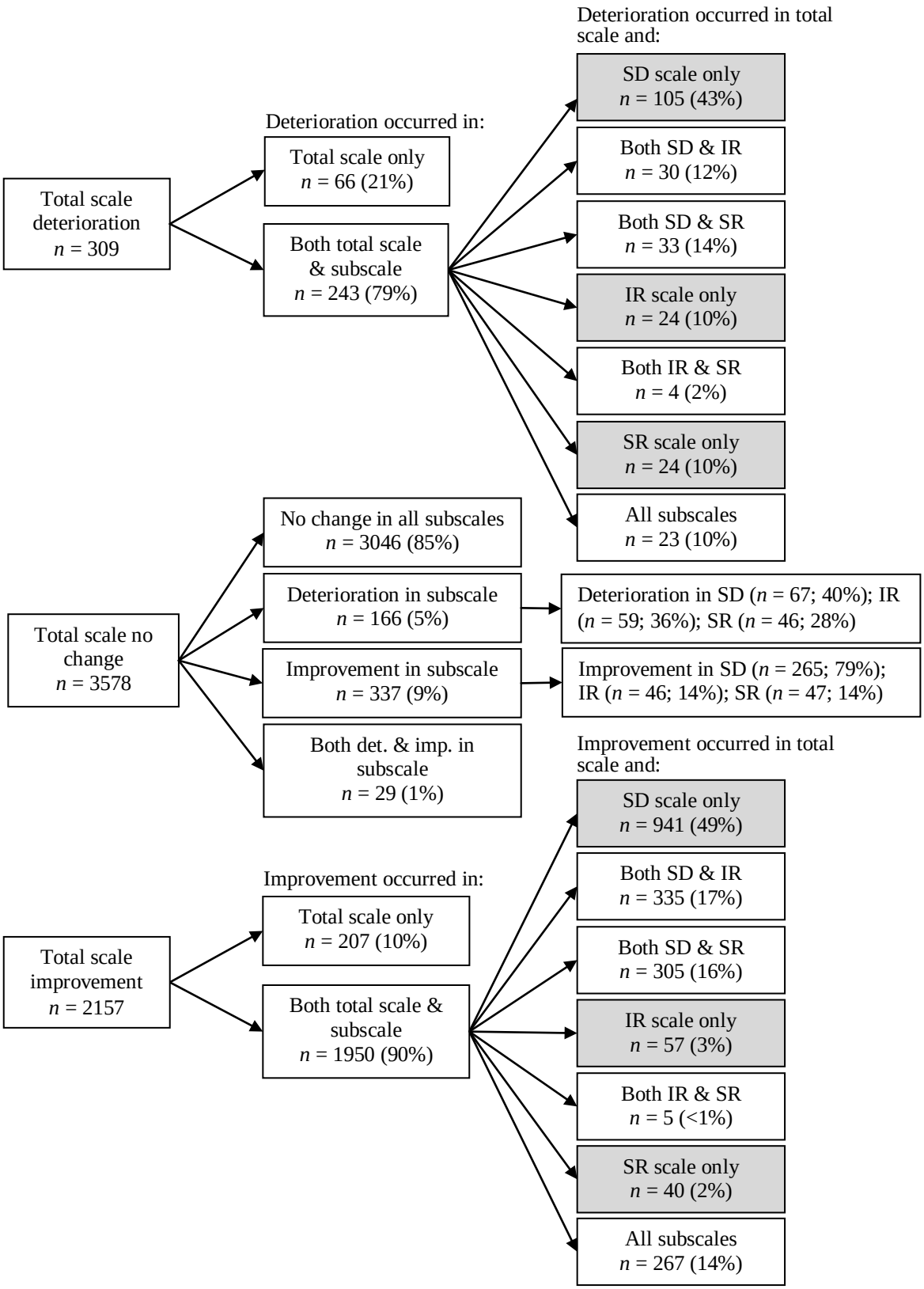


Figure 1.

Breakdown of clinical significance match between the total scale and subscales.

Appendix B

Calculation of Effect Sizes and Standard Errors

Effect Sizes: Post Treatment Score Comparison

Hedge's standardized mean difference (g) for mean post-test OQ scores and mean session attendance comparisons is calculated as the following (Comprehensive Meta-Analysis Version 2; Hedges, 1981):

1. Calculate standardized difference in means (d) by dividing the raw score difference in means by pooled standard deviation of two samples ($M_1 - M_2$) in comparison: $d = (M_1 - M_2)/s_{pooled}$ where s_{pooled} is calculated by using the following formula:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)\times s_1^2 + (n_2 - 1)\times s_2^2}{(n_1 + n_2 - 2)}}. \quad (1)$$

where n_1 and n_2 represent the sample sizes of samples 1 and 2, and s_1 and s_2 represent the standard deviations of samples 1 and 2.

2. Compute correction factor J for correcting bias: $J = 1 - [3/(4df - 1)]$, where df is given by

$$df = n_1 + n_2 - 2. \quad (2)$$

3. Compute Hedge's standardized mean difference (g) by multiplying d by a correction factor (J): $g = d \times J$. (3)

Standard Errors

1. Obtain standard error for standard difference in means (d):

$$SE(d) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{2(n_1 + n_2)}}. \quad (4)$$

2. Correct for bias by multiplying standard error of standardized mean difference by a correction factor J:

$$SE(g) = SE(d) \times J. \quad (5)$$

Effect Sizes: Pre-Post Change Score per Group

1. Obtain independent standardized mean difference between pre-treatment and post-treatment ($d_{pre\text{-}post}$) by:

$$d_{pre-post} = \frac{(M_{pre} - M_{post})}{(\frac{s_{pre-post}}{\sqrt{2\,(1-r_{pre-post})}})} \quad (6)$$

where ($M_{pre} - M_{post}$) represents the difference in the mean pre-treatment scores and the mean post-treatment scores, $r_{pre\text{-}post}$ represents the correlation coefficient between the pre-treatment and post-treatment scores, and $s_{pre\text{-}post}$ is obtained by

$$s_{pre-post} = \sqrt{s_{pre}^2 + s_{post}^2 - 2 \times r_{pre-post} \times s_{pre} \times s_{post}} \quad (7)$$

2. Calculate a correction factor *J* by following equation (2) and apply this correction factor to *dpre-post* as obtained by equation (6):

$$g_{pre-post} = d_{pre-post} \times J\,. \quad (8)$$

Standard Errors for Pre-Post Change Effect Size

1. Obtain standard error for independent standardized mean difference between pre-treatment and post-treatment by:

$$SE_{d_{pre-post}} = \frac{s_{pre-post}}{\sqrt{n}} \quad (9)$$

2. Correct for bias by applying a correction factor *J*:

$$SE_{g_{pre-post}} = SE_{d_{pre-post}} \times J\,. \quad (10)$$

Effect Sizes: Pre-Post Change Score Comparison

Effect size for difference in pre-post change scores between two groups is obtained in the similar manner as for the standard mean difference in post-treatment scores as described above. In place of formula (1), the following formulae are used.

Calculate standardized difference in mean change scores (d_{change}) by dividing the raw score difference in mean change scores ($M_{change\,1} - M_{change\,2}$) by pooled standard deviation of two samples in comparison:

$$d_{change} = (M_{change\,1} - M_{change\,2})/s_{change\_pooled} \quad (11)$$

where s_{change_pooled} is given by

$$s_{change\_pooled} = \sqrt{\frac{(n_1 - 1)\times s^2_{change\,1} + (n_2 - 1)\times s^2_{change\,2}}{(n_1 + n_2 - 2)}} \quad (12)$$

and $s_{change\,1}$ and $s_{change\,2}$ are given by

$$s_{change\,1} = \sqrt{s^2_{pre\,1} + s^2_{post\,1} - 2 \times r_{pre-post} \times s_{pre\,1} \times s_{post\,1}} \quad (13)$$

and

$$s_{change\,2} = \sqrt{s^2_{pre\,2} + s^2_{post\,2} - 2 \times r_{pre-post} \times s_{pre\,2} \times s_{post\,2}} \; . \quad (14)$$

In equations (13) and (14), $r_{pre-post}$ is the weighted mean correlation coefficient between pre-treatment and post-treatment scores of sample 1 and sample 2. $s_{pre\,1}$ and $s_{pre\,2}$ represent the pre-treatment standard deviations of samples 1 and 2, and $s_{post\,1}$ and $s_{post\,2}$ represent the post-treatment standard deviations of samples 1 and 2.

Appendix C

Assignment of Random Weight

Random weight and the main effect are calculated as the following (Hedges & Vevea, 1998; Comprehensive Meta-Analysis, Version 2):

A random weight (*w*) assigned to each individual study (*i*) is defined as:

$$w_i = \frac{1}{v_i^*} \tag{15}$$

where v_i^* represents the sum of within-study variance study (*i*) and the between-studies variance (τ^2):

$$v_i^* = v_i + \tau^2. \tag{16}$$

The mean effect size ($\bar{g}$) is calculated as:

$$\bar{g} = \frac{\sum_{i=1}^{k} w_i g_i}{\sum_{i=1}^{k} w_i}. \tag{17}$$

The variance of the mean effect is defined as the reciprocal of sum of the individual study weights. Thus, the standard error (*SE*) of the mean effect is the square root of the sampling variance:

$$SE(\bar{g}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}. \tag{18}$$

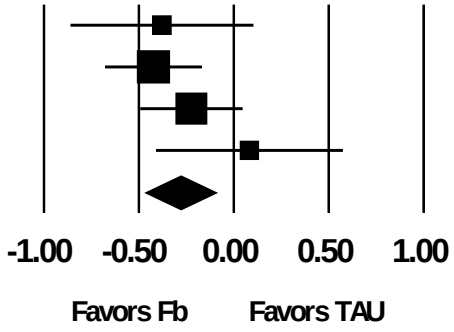
Appendix D

Detailed Results of Meta-analyses and Forest Plots

Table 1

ITT Analysis of Post-Treatment OQ Score Comparisons (*Fb vs. TAU)

Study name	Hedges's g	Standard error	p-Value
Lambert et al., 2001	-0.38	0.25	0.124
Lambert et al., 2002	-0.42	0.13	0.001
Whipple et al., 2003	-0.22	0.14	0.106
Hawkins et al., 2004	0.08	0.25	0.743
	-0.28	0.09	0.003

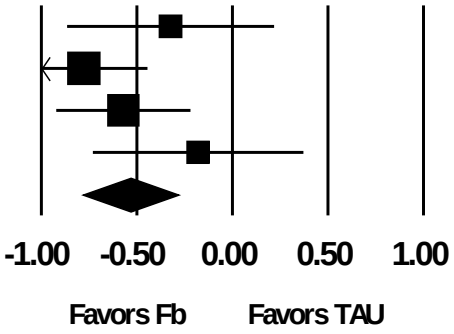


*Fb = Feedback group; TAU = Treatment as usual

Table 2

Efficacy Analysis of Post-Treatment OQ Score Comparisons (*Fb vs. TAU)

Study name	Hedges's g	Standard error	p-Value
Lambert et al., 2001	-0.32	0.28	0.242
Lambert et al., 2002	-0.78	0.17	0.000
Whipple et al., 2003	-0.57	0.18	0.001
Hawkins et al., 2004	-0.18	0.28	0.523
	-0.53	0.13	0.000

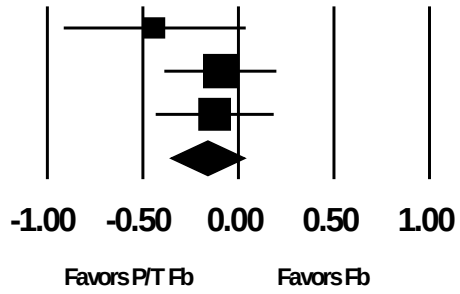


*Fb = Feedback group; TAU = Treatment as usual

Table 3

ITT Analysis of Post-Treatment OQ Score Comparisons (*P/T Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Hawkins et al., 2004	-0.44	0.24	0.071
Harmon et al., 2007	-0.09	0.15	0.526
Slade et al, 2008	-0.12	0.16	0.430
	-0.16	0.10	0.099

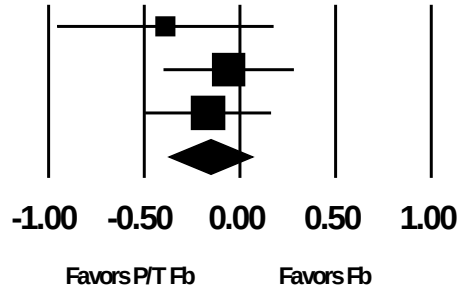


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 4

Efficacy Analysis of Post-Treatment OQ Score Comparisons (*P/T Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Hawkins et al., 2004	-0.39	0.29	0.177
Harmon et al., 2007	-0.06	0.17	0.734
Slade et al, 2008	-0.17	0.17	0.322
	-0.16	0.11	0.163

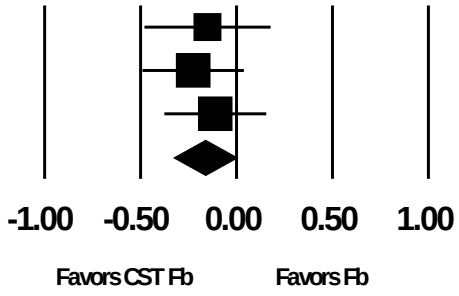


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 5

ITT Analysis of Post-Treatment OQ Score Comparisons (*CST Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Whipple et al., 2003	-0.15	0.17	0.368
Harmon et al., 2007	-0.23	0.13	0.094
Slade et al, 2008	-0.11	0.14	0.415
	-0.16	0.08	0.048

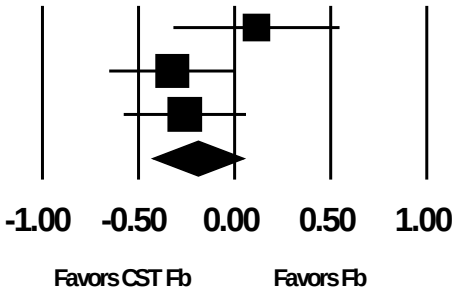


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 6

Efficacy Analysis of Post-Treatment OQ Score Comparisons (*CST Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Whipple et al., 2003	0.11	0.22	0.606
Harmon et al., 2007	-0.32	0.17	0.053
Slade et al, 2008	-0.26	0.16	0.110
	-0.19	0.12	0.113

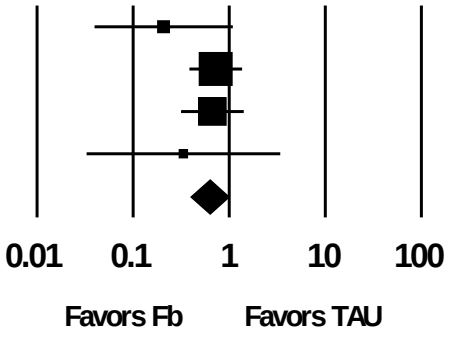


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 7

ITT Analysis of Odds of Clinically Significant Worsening (*Fb vs. TAU)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Lambert et al., 2001	0.21	0.04	1.09	0.063
Lambert et al., 2002	0.72	0.39	1.36	0.315
Whipple et al., 2003	0.67	0.32	1.42	0.294
Hawkins et al., 2004	0.33	0.03	3.40	0.354
	0.62	0.40	0.98	0.040

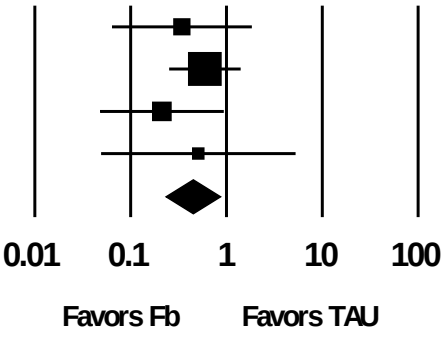


*Fb = Feedback group; TAU = Treatment as usual

Table 8

Efficacy Analysis of Odds of Clinically Significant Worsening (*Fb vs. TAU)

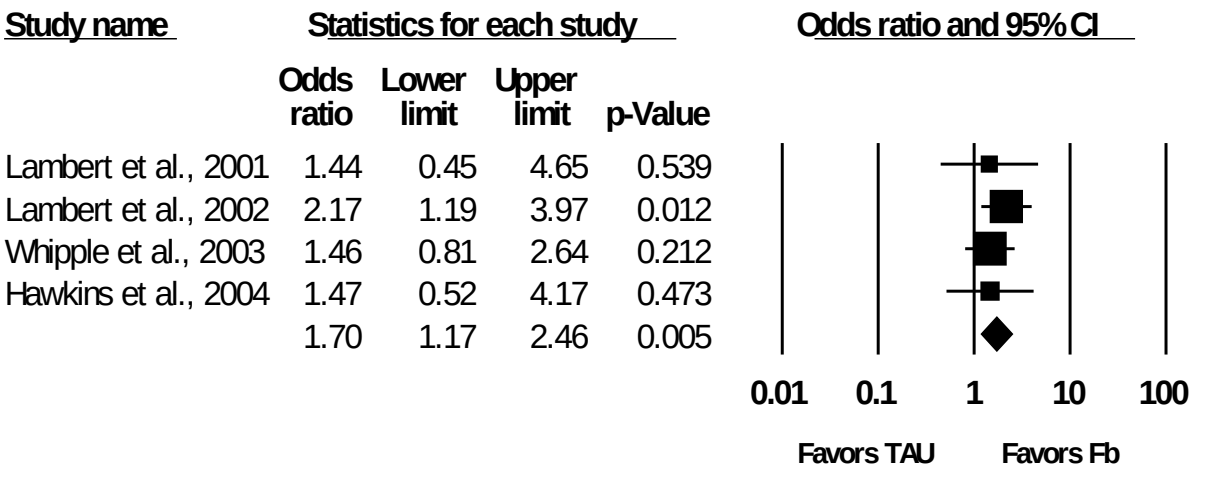
Study name	Odds ratio	Lower limit	Upper limit	p-Value
Lambert et al., 2001	0.34	0.06	1.84	0.212
Lambert et al., 2002	0.60	0.25	1.41	0.238
Whipple et al., 2003	0.21	0.05	0.94	0.041
Hawkins et al., 2004	0.51	0.05	5.26	0.571
	0.44	0.23	0.85	0.015



*Fb = Feedback group; TAU = Treatment as usual

Table 9

ITT Analysis of Odds of Clinically Significant Improvement (*Fb vs. TAU)

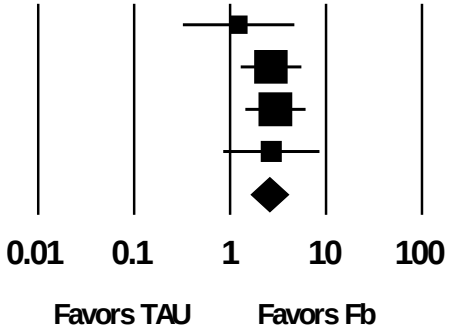


*Fb = Feedback group; TAU = Treatment as usual

Table 10

Efficacy Analysis of Odds of Clinically Significant Improvement (*Fb vs. TAU)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Lambert et al., 2001	1.23	0.32	4.67	0.766
Lambert et al., 2002	2.67	1.29	5.53	0.008
Whipple et al., 2003	2.97	1.44	6.11	0.003
Hawkins et al., 2004	2.69	0.85	8.54	0.093
	2.55	1.64	3.98	0.000

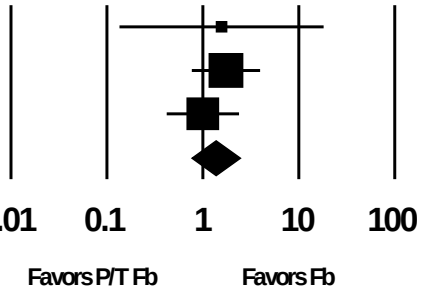


*Fb = Feedback group; TAU = Treatment as usual

Table 11

ITT Analysis of Odds of Clinically Significant Worsening (*P/T Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Hawkins et al., 2004	1.57	0.14	18.15	0.719
Harmon et al., 2007	1.74	0.77	3.95	0.184
Slade et al, 2008	1.00	0.42	2.38	1.000
	1.35	0.76	2.41	0.306

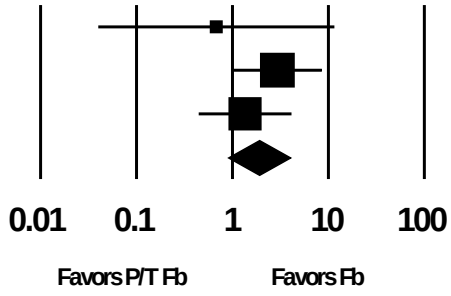


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 12

Efficacy Analysis of Odds of Clinically Significant Worsening (*P/T Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Hawkins et al., 2004	0.68	0.04	11.53	0.788
Harmon et al., 2007	2.95	1.02	8.54	0.047
Slade et al, 2008	1.35	0.44	4.13	0.594
	1.89	0.90	3.96	0.094

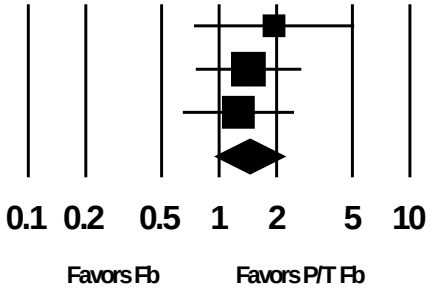


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 13

ITT Analysis of Odds of Clinically Significant Improvement (*P/T Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Hawkins et al., 2004	1.94	0.74	5.10	0.179
Harmon et al., 2007	1.43	0.76	2.70	0.274
Slade et al, 2008	1.26	0.65	2.47	0.495
	1.44	0.95	2.19	0.086

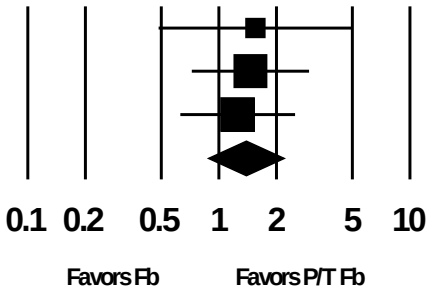


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 14

Efficacy Analysis of Odds of Clinically Significant Improvement (*P/T Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Hawkins et al., 2004	1.55	0.48	4.99	0.459
Harmon et al., 2007	1.46	0.72	2.96	0.291
Slade et al, 2008	1.25	0.63	2.50	0.521
	1.38	0.88	2.18	0.164

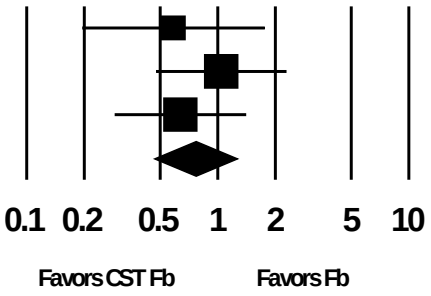


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 15

ITT Analysis of Odds of Clinically Significant Worsening (*CST Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Whipple et al., 2003	0.59	0.20	1.76	0.342
Harmon et al., 2007	1.04	0.48	2.29	0.916
Slade et al, 2008	0.64	0.29	1.41	0.266
	0.76	0.46	1.26	0.288

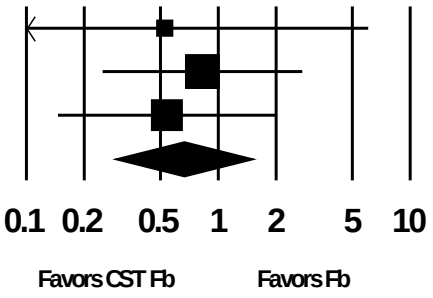


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 16

Efficacy Analysis of Odds of Clinically Significant Worsening (*CST Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Whipple et al., 2003	0.53	0.05	6.05	0.606
Harmon et al., 2007	0.83	0.25	2.74	0.756
Slade et al, 2008	0.54	0.15	2.00	0.356
	0.66	0.29	1.52	0.329

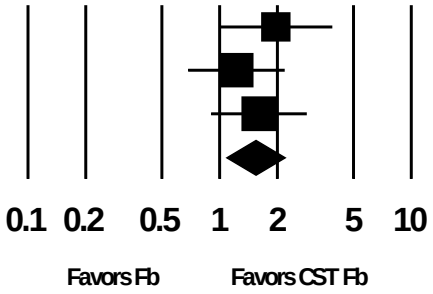


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 17

ITT Analysis of Odds of Clinically Significant Improvement (*CST Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Whipple et al., 2003	1.97	1.00	3.87	0.050
Harmon et al., 2007	1.22	0.68	2.18	0.497
Slade et al, 2008	1.60	0.90	2.85	0.107
	1.53	1.08	2.18	0.016

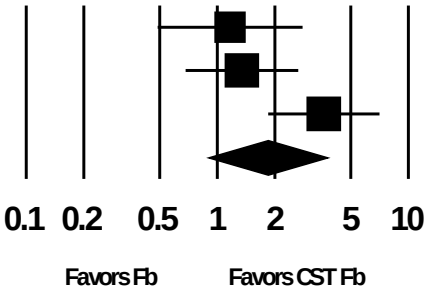


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 18

Efficacy Analysis of Odds of Clinically Significant Improvement (*CST Fb vs. Fb)

Study name	Odds ratio	Lower limit	Upper limit	p-Value
Whipple et al., 2003	1.17	0.49	2.79	0.729
Harmon et al., 2007	1.35	0.68	2.65	0.392
Slade et al, 2008	3.61	1.85	7.06	0.000
	1.83	0.89	3.76	0.098

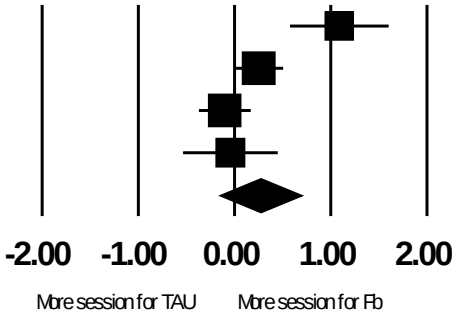


*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group

Table 19

ITT Analysis of Mean Session Attendance (* Fb vs. TAU)

Study name	Hedges's g	Standard error	p-Value
Lambert et al., 2001	1.09	0.26	0.000
Lambert et al., 2002	0.25	0.13	0.052
Whipple et al., 2003	-0.10	0.14	0.459
Hawkins et al., 2004	-0.04	0.25	0.865
	0.27	0.22	0.217

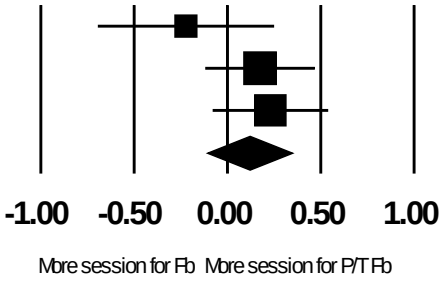


*Fb = Feedback group; TAU = Treatment as usual

Table 20

ITT Analysis of Mean Session Attendance (*P/T Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Hawkins et al., 2004	-0.22	0.24	0.356
Harmon et al., 2007	0.17	0.15	0.244
Slade et al, 2008	0.23	0.16	0.145
	0.12	0.12	0.311

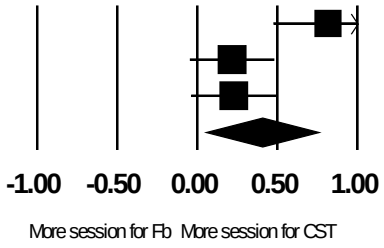


*P/T Fb = Patient and therapist feedback group; Fb = Feedback group

Table 21

ITT Analysis of Mean Session Attendance (*CST Fb vs. Fb)

Study name	Hedges's g	Standard error	p-Value
Whipple et al., 2003	0.82	0.17	0.000
Harmon et al., 2007	0.22	0.13	0.106
Slade et al, 2008	0.23	0.14	0.093
	0.40	0.18	0.024



*CST Fb = Clinical Support Tools feedback group; Fb = Feedback group