2010-07-16

Change Trajectories and Early Warning System to Identify Youth at Risk for Negative Psychotherapy Outcome

Philip Legrand Nelson
Brigham Young University - Provo

Change Trajectories and Early Warning System to Identify Youth at Risk for Negative Psychotherapy Outcome

Philip L. Nelson

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

Doctor of Philosophy

Robert L. Gleave, Chair
Gary M. Burlingame
Jared S. Warren
Ellie L. Young
Lane Fischer

Department of Counseling Psychology and Special Education

Brigham Young University

August 2010

ABSTRACT

Change Trajectories and Early Warning System to Identify

Youth at Risk for Negative Psychotherapy Outcome

Philip L. Nelson

Department of Counseling Psychology and Special Education

Doctor of Philosophy

The field of mental health treatment is making efforts to better serve all psychotherapy clients, but especially the 5–10% of clients who deteriorate in treatment (Lambert & Ogles, 2004) and the 30–60% who drop out prematurely (Pekarik & Stephenson, 1988). These efforts involve collaboration between research and practice. Both research and practice have been treatment focused for much of their history, primarily examining treatment efficacy or effectiveness, and never quite settling on the generalizability or applicability of specific treatments. The patient-focused research paradigm has shifted the focus from treatment outcomes on the group level to outcomes on the individual client level. This movement involves outcome monitoring for purposes of treatment planning and quality care. Some of these monitoring systems include early warning systems that could help identify and better serve clients who are at risk for negative outcome.

The present study validated previous warning system studies for youth and replicated tests for variables that were predictive of youth change trajectories using the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004). This study also replicated the accuracy of a warning system for at-risk youth clients, exploring various approaches to creating the cutoffs the warning system uses for its predictions, and reporting the respective accuracy of each. This study contributes to future studies comparing outcomes between client groups whose therapists do or do not receive systematic feedback. This endeavor offers many benefits to quality improvement efforts being made by clinicians and managed care organizations.

Keywords: warning system, psychotherapy outcomes, youth, change trajectories

ACKNOWLEDGEMENTS

Robert Gleave has been a life mentor for me, modeling wisdom and discretion, and offering unfaltering support and patience. He has demonstrated a respect for and confidence in me that I cherish. Close friends and family have sustained me and have helped me identify important nonacademic lessons in an academic experience. Jared Warren has been a patient support and perseverant collaborator, greatly boosting my academic output. Gary Burlingame has also been an excellent resource and mentor for research and general productivity, sharing rare opportunities and access to a superb research network. I'm grateful and indebted to these people. I am also grateful to the many other key individuals in the Counseling Psychology program, and the broader university, whose investment exceeded their obligation in providing resources, support, and context for my graduate experience. The experience has provided me a priceless and transformative deepening of emotion, intellect, and spirituality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

People entering psychotherapy may appropriately hope for positive outcomes because psychotherapy is effective for most clients (Grissom, 1996; Lambert & Ogles, 2004; Lipsey & Wilson, 1993; Shapiro & Shapiro, 1982; Smith, Glass, & Miller, 1980). However, this hope goes unrealized for a number of clients whose symptoms do not improve. Symptoms for 5–10% of clients are worse after treatment than before (Bishop et al., 2005; Lambert & Bergin, 1994; Lambert & Ogles, 2004; Mohr, 1995; Weisz, Donenberg, Han, & Weiss, 1995) and 30–60% of clients drop out of treatment early (Pekarik & Stephenson, 1988).

Ideally, clinicians would quickly identify and attend to clients at risk for negative outcome, but on their own, clinicians identify as few as 2.5% of deteriorators (Hannan et al., 2005). Even though their own prediction accuracy is lower than that of empirical methods, clinicians are often reluctant to trust research-based methods for identifying at-risk clients (Grove & Meehl, 1996). Their reluctance typically concerns the extent to which research findings from highly controlled experimental settings can truly apply to real-world clinical practice. This is typical of a widespread divide between research and practice. Fortunately the divide is shrinking as researchers and practitioners collaborate to focus on client care (Kazdin, 2008). Some collaborations have focused on creating early warning systems to identify clients at risk for negative outcome (e.g., Finch, Lambert, & Schaalje, 2001; Harmon et al., 2007; Lambert, Hansen, & Finch, 2001).

Unfortunately, the development of early warning systems for youth clients is only just gaining momentum (e.g., Bishop et al., 2005; Bybee, Lambert, & Eggett, 2007; Cannon, Warren, Nelson, & Burlingame, 2010; Warren, Nelson, Mondragon, Baldwin, & Burlingame, 2010). There remains a dire need for outcome monitoring and early warning systems for youth (Burns,

Hoagwood, & Mrazek, 1999; Weisz & Gray, 2008; Weisz, Jensen, & McLeod, 2005). Highlighting this need are treatment effect sizes near zero for youth in some settings (Weisz, 2004; Weisz, Donnenberg, et al., 1995), estimates suggesting that 40–60% of youth drop out of treatment early (Kazdin, 1996; Wierzbicki & Pekarik, 1993), and more than 10% of youth whose symptoms are worse after treatment than before (Cannon et al., 2010; Kazdin, 2003; Shirk & Russell, 1992; Weisz, Donnenberg, et al., 1995).

The mental health research literature has not fully investigated the composition and administration of treatments for children and adolescents, nor does it fully understand typical patterns of change in response to psychotherapy treatments (Garland, Hurlburt, & Hawley, 2006; Kazdin, 2000). In brief, the youth literature lacks studies performed in real world settings (Weisz et al., 2005). Considering the millions of youth in psychotherapy treatment each year (National Advisory Mental Health Council, 2001; Ringel & Sturm, 2001), non-responders constitute a rather large number of children and adolescents. Action must be taken to shift youth non-responders' treatment experience from false hope to legitimate help.

The present study takes an important next step in the development of outcome monitoring and early warning systems for youth by validating previous studies and replicating tests for variables that are predictive of youth change trajectories. This study also replicated the accuracy of a warning system for at-risk youth clients, using the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004). The results from this study contribute to the understanding and application of warning systems to clinical settings for youth. This sets the stage for future studies comparing outcomes between client groups whose therapists do or do not receive systematic feedback. This effort offers many benefits to quality improvement efforts by clinicians and managed care organizations.

LITERATURE REVIEW

Psychotherapy researchers and practitioners have struggled to fully integrate their efforts to explore and improve psychotherapy. Their contextual differences call into question how well findings generalize between their respective settings. However, patient-focused research (Howard, Moras, Brill, Martinovich, & Lutz, 1996) circumvents some problems of generalizability by facilitating individualized outcome monitoring and treatment modification for psychotherapy clients. Early warning systems assist in such ongoing evaluation of outcomes, drawing clinicians' attention to clients at risk for negative outcomes (e.g., Finch et al., 2001; Harmon et al., 2007; Lambert et al., 2001). Although early warning systems have been associated with improved outcomes for adult psychotherapy clients, such systems are not as fully developed for youth client populations. The present study contributes to the research literature regarding predictors of youth outcomes in psychotherapy. It also replicated the accuracy of an early warning system for at-risk youth clients, using the YOQ (Burlingame et al., 2004).

Struggles to Integrate Psychotherapy Research and Practice

Roughly 10% of psychotherapy clients experience negative outcomes and even more experience no clinically significant response to treatment (Bishop et al., 2005; Lambert & Bergin, 1994; Lambert & Ogles, 2004; Mohr, 1995; Weisz, Donenberg, et al., 1995). Ideally, psychotherapists would quickly identify and attend to these at-risk clients. Unfortunately, therapist judgment of expected outcomes is poor. Even though research-based identification methods are rather accurate, clinicians commonly resist using them because of concerns regarding the applicability of research in real-world clinical practice (Grove & Meehl, 1996). These concerns over applicability are well founded, considering the history and nature of psychotherapy research. The next section explores therapists' accuracy in predicting client

outcome. The following sections explore two major movements in psychotherapy research—efficacy and effectiveness research—and threats to their applicability in clinical practice.

Therapists' Predictions of Client Outcomes

Therapist judgment of expected outcomes is poor (Grove & Meehl, 1996), even for therapists with ample clinical experience (Dawes, 1989). Hannan et al. (2005) replicated the finding of many other studies (Grove & Meehl, 1996), demonstrating therapists' inferior prediction of client outcome in comparison with empirically derived systems. Despite being informed of the 8% deterioration rate for their clinic, the 48 therapists participating in the study predicted that only 3 of 550 clients (0.01%) would deteriorate. In actuality, 40 clients (7.3%) ended up deteriorating, only one of which had been identified by the therapists. Thus clinicians identified 2.5% of deteriorators and the warning system identified 86% (by the third session). This study suggests that therapists' outcome predictions may be overly optimistic and far less accurate than research-based warning systems.

Beyond the issue of poor prediction accuracy, therapists commonly have the misconception that clients' conditions worsen before improving (Canen & Lambert, 1999), perhaps as the clients more fully confront and realize the extent of their challenges. Some therapists encourage new clients to persevere through the initial discomfort of gaining momentum in treatment, but perhaps these therapists' attention to a possible heightening of symptoms has led them to expect it as typical, rather than indicative of ineffective treatment. In actuality, early deterioration is a risk factor for deterioration as a final outcome (Haas, Hill, Lambert, & Morrell, 2002). On the other hand, gains in early treatment are common (Wilson, 1999) and are among the best predictors of positive final outcomes (Haas et al., 2002).

Considering therapists' poor prediction accuracy, their misconceptions regarding outcome predictors, and the superior predictions of research-based warning systems, why do therapists trust their clinical judgment more than research? Grove and Meehl (1996) were somewhat unforgiving in their review and rebuttal of clinicians' many arguments against incorporating research results into clinical practice. Kazdin (2008) balanced the arguments somewhat for the research–practice debate, exploring ways to find unity. His review explains the goals and shortcomings of two major movements in research and practice, representing the efficacy movement in terms of evidence-based treatments (EBTs) and representing the effectiveness movement in terms of evidence-based practices (EBPs). The next sections review how these two movements fostered and maintained the divide between research and practice. A later section explores the potential of a third movement—patient-focused research—to shrink the gap between research and practice.

Efficacy Research: Evidence-based Treatments

Efficacy research has been the mainstay of quantitative research in psychotherapy treatment. It is typified by randomized clinical trials comparing experimental treatment groups to criterion or control groups. It uses rigorous experimental control of potential covariates and confounds in attempt to ensure that observed effects are truly attributable to the experimental treatment (Howard, et al., 1996). Treatments demonstrating efficacy on the aggregate level gain the status of evidence-based treatments (EBTs; Kazdin, 2008).

The tight controls that offer efficacy research its internal validity are the very attributes that threaten its external validity (i.e., generalizability) and are the target of practitioners' complaints regarding applicability (Chambless & Hollon, 1998; Cook & Campbell, 1979; Howard et al., 1996; Kazdin, 2008). Randomized assignment to experimental or control groups

attempts to avoid systematic differences between groups that could confound the treatment results. Although group assignment may be random, attrition (i.e., dropout) typically is not, thus jeopardizing the ability of randomization to reach its goals of ensuring sample comparability (Howard, Krause, & Lyons, 1993). Although larger samples may remain fairly immune to problematic attrition, treatment effects observed in smaller and more susceptible samples must be replicated by additional studies (Howard, Kopta, & Orlinsky, 1986).

Other study controls such as stringent inclusion criteria (e.g., clients with specific single diagnoses, specific demographics, etc.) and manualized treatments attempt to reduce heterogeneity in research conditions that might yield error, or at least create "noise" in the study's results. However, clinicians (and many researchers) complain that such homogeneous study conditions produce results that are not generalizable to clinical practice, which typically has heterogeneous conditions (e.g., variety of client demographics, comorbid diagnoses, etc.; Goldfried & Wolfe, 1998; Seligman, 1995).

With careful research design ensuring internal validity, most clinicians agree that EBTs work. However, these clinicians add a qualifier: "EBTs work *in the experimental setting*," and may add the question, "…but do they work in the clinical setting, in my setting?" This becomes the question of effectiveness research, a movement that attempts to maximize external validity and generalizability. The next section provides a review of effectiveness research.

Effectiveness Research: Evidence-based Practices

Effectiveness research attempts to remediate the generalizability concerns of efficacy research by performing studies in naturalistic or real-world clinical settings. These studies attempt to identify treatments that work in actual clinical practice and in light of clients'

heterogeneity and individual differences (Chambless & Hollon, 1998). Treatments that work on an aggregate level gain the status of evidence-based practices (EBPs; Kazdin, 2008).

Unfortunately, the naturalistic research design that offers effectiveness research its external validity also includes a number of threats to internal validity. The primary threats are the lack of experimental controls. Treatment group assignment can rarely be random and the effects of unobserved/unmeasured variables often remain unknown, which throws into question the appropriateness of attributing effects to experimental treatments. Study results must be interpreted with caution and conclusions regarding treatment effectiveness typically require a number of replication or validation studies (Howard et al., 1986).

With keen insight, Kazdin (2008) raised the philosophical argument that effectiveness research may in fact have low generalizability, despite its intentions to the contrary. For example, if clients are so unique and individual differences have such bearing on treatment effectiveness—as is the fundamental concern driving the advent of effectiveness research in response to efficacy research—the more than 32,000 symptom combinations meeting criteria for a diagnosis of conduct disorder (demographic variables omitted) must seriously threaten the likelihood that a treatment successful with one individual's set of symptoms would generalize to and be successful with symptoms for another individual (Perepletchikova & Kazdin, 2005).

Need for a Paradigm Bridging the Gap between Efficacy and Effectiveness

Efficacy and effectiveness research each have advantages and disadvantages. Their disparity is at the heart of the gap between research and practice. The generalizability of efficacy research is admittedly questionable, but effectiveness research does not necessarily appear to offer an infallible solution. Given the background presented above, generalizability may be limited for both efficacy and effectiveness research.

An additional shared weakness with efficacy and effectiveness research is that they do little to address the issue of non-responders and deteriorators. Instead, they are both treatment focused, concerning themselves only with how treatments function on the aggregate level (Howard et al., 1996). Although these research paradigms identify treatments that work for specific populations, clients' individual characteristics may nonetheless influence their therapy experience and outcomes (Huffman, Martin, Botcheva, Williams, & Dyer-Friedman, 2004). Using treatments based in either research paradigm, how might a therapist respond to individual client complaints of non-improvement? A tempting, but likely inappropriate response from the therapist might be, "I'm sorry that you're not getting better. We only use treatment types and delivery styles shown to be the best for most people. This is the best we can do." On the contrary, a new patient-focused research paradigm helps therapists do better than this with clients who appear unresponsive to treatment (Howard et al., 1996).

Efforts to help non-responders need not abandon efficacy and effectiveness research. Both paradigms are valuable and have made great contributions to the field of mental health treatment. However, some rapprochement between the two is necessary to improve the quality of client care. Systematic outcome monitoring to ensure quality of care for each client is one example of rapprochement (Kazdin, 2008). Offering EBTs and EBPs is a great start, but patient-focused research goes a step further to evaluate what works for a given client in a given context, making adjustments and accommodations throughout treatment. Research in this area has begun developing systems for ongoing evaluation of individual clients' progress, providing therapists with real-time feedback (Brown, Lambert, Jones, & Minami, 2004; Cattani-Thompson, 2003; Finch et al., 2001). The next section explores patient-focused research and will lead into an

examination of how this research can serve as the foundation of early warning systems that help clinicians quickly identify clients who are at risk for negative outcome.

Patient-focused Research

Research with EBTs and EBPs alone will likely never fully remediate the problem that some clients do not improve along with the majority, the obstacle being that these research paradigms are treatment focused and only examine the group level, without attention to aberrant individuals (Howard et al., 1996). Clinicians alone may not be able to solve the non-responder problem either, their obstacle being their poor accuracy predicting which clients will experience negative outcome (Grove & Meehl, 1996). Instead, researchers and clinicians uniting to focus on quality care for individual clients may have the most potential to help individual clients whose treatment appears ineffective (Kazdin, 2008). This is a central aim of patient-focused research, which uses outcome measures to monitor and adjust treatment for individual clients (Lutz, Martinovich, Howard, & Leon, 2002).

In an effort similar to the patient-focused movement, the American Psychological Association (APA) created a task force for evidence-based practice in psychology (EBPP; APA, 2006). Their purpose was "integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences" (p. 273). They acknowledged that therapists are generally biased in their judgments and suggested that outcome monitoring and feedback be used to counteract such biases. They proposed that therapists monitor individual outcomes and adjust treatment as appropriate, as opposed to merely examining average group responses to treatments.

Outcome monitoring has a number of potential benefits for research and practice. Change trajectories plotting scores over time provide insight for the process of change in terms of

magnitude and timing. For example, rapid initial change may indicate more frequent outcome monitoring in early treatment. Researchers can explore change trajectories specific to treatment concern and intervention type, potentially informing or testing theory (Ilardi & Craighead, 1999; Laurenceau, Hayes, & Feldman, 2007; Tang & DeRubeis, 1999a, 1999b).

Among the greatest benefits of patient-focused research is its potential to facilitate ongoing treatment evaluation and ultimately serve as the foundation for an early warning system to identify clients at risk for negative outcome. Therapist feedback studies may be too scarce (Davis, Thompson, Oxman, & Haynes, 1995) to remediate clinicians' inaccurate judgments regarding their clients' eventual outcomes (Claiborn & Goodyear, 2005; Hannan et al., 2005). In addition, it appears that clinicians have had difficulty incorporating feedback into their judgments of client progress (Nisbett & Ross, 1980; Rossi, Schuerman, & Budde, 1996), perhaps because the feedback has been too global or has arrived too late to be useful (Garb & Shramke, 1996). Outcome monitoring data are available in many settings, but it has been challenging to formulate and provide feedback to therapists in a timely and effective manner (Lambert, Hansen, et al., 2001; Saptya, Reiman, & Bickman, 2004). The following section reviews a number of existing outcome monitoring systems, describing how they formulate and deliver feedback to therapists.

Early Warning Systems Predicting Negative Psychotherapy Outcomes

As a product of patient-focused research, early warning systems have potential to address the problems of premature dropout and negative outcome among psychotherapy clients. Effective systems warn therapists regarding clients who are not progressing as expected or who are following a path typical of those who deteriorate or drop out of treatment early. For warning systems to detect such deviations from normal progress in treatment, they must track actual

outcomes using a reliable and valid outcome measure. Ideally, this measure is sensitive to change in clients' symptoms and remains valid during repeated administrations.

Warning systems typically have systematic criteria for what deviation identifies clients as at risk for negative outcome. These criteria occasionally compare clients' ongoing outcome to their personal baselines, but other times compare ongoing outcome with expected outcome. The outcomes many systems expect are simply the mean outcomes observed in actual clients, calculated using descriptive or inferential statistics. Some warning systems use expected outcomes that differ by client subpopulation, each subpopulation sharing particular characteristics (e.g., initial severity, sex, and other demographics). The sections that follow present some existing early warning systems. The early warning function is often only one component of broader and more fully developed outcome monitoring systems that aid clinicians' judgment of clients' current functioning (i.e., clinical or nonclinical range), current trajectory (i.e., on track, not responding, deteriorating) and likely final outcome.

Warning Systems in Development

The several warning systems described below are apparently still in development or their detailed information appears to be inaccessible. One tracks outcome but lacks an algorithm for alerting clinicians to clients at risk for negative outcome. The others lack information about their prediction accuracy. The descriptions of each system mention the system's outcome measures, criteria for ongoing outcomes that identify clients as at-risk, and method of generating comparative expected outcomes, if any.

Systematic Treatment Selection. Fisher, Beutler, and Williams (1999) described Systematic Treatment Selection, a procedure of matching client symptoms to specific treatments, and matching clients and treatments to specific therapists. The system's matching procedure is

intended to improve therapy outcomes. The system also includes an outcome tracking component to aid treatment planning and quality of care. Fisher and colleagues indicated that with further developments the system could alert clinicians to clients at risk for negative outcome. The system relies primarily on therapists' ratings of client outcomes in attempt to avoid unreliable self-reporting from clients. Considering the demand this puts on therapists, along with the highly computerized nature of this system, the Systematic Treatment Selection procedure may not be very feasible for widespread implementation as an early warning system.

Stuttgart-Heidelberg model. The Center for Psychotherapy Research Stuttgart and the Psychiatric Clinic of the University of Heidelberg collaborated to create an outcome monitoring system they called the Stuttgart-Heidelberg model (Kordy, Hannover, & Richard, 2001). This model shifts away from intrusive quality assurance programs to a bottom-up approach that prioritizes problem detection and problem solution rather than institutional sanctions. It attempts to ensure quality of outcome rather than just quality of structure and of process (Donabedian, 1982). The creators' viewpoint was that treatment failures are significant and deserve attention and prevention.

For outcome tracking, the Stuttgart-Heidelberg model uses periodic administration of the Severity of Impairment Score (BSS; Schepank, 1995), along with additional measures specific to the treatment concerns and context. The BSS can be completed using a computer or using paper and pencil. The system identifies clients as at risk for negative outcome if ever their scores surpass an "action limit." Kordy et al. (2001) provide little information on what this action limit is and how it is derived. It appears that scores crossing this threshold demonstrate a dangerous level of deterioration, as though the threshold is the boundary on one side of a confidence

interval around scores. The model also has a reliable change index (RCI; Jacobson & Truax, 1991) to classify final outcomes as reliably changed for better or worse.

The Stuttgart-Heidelberg model provides therapists three levels of feedback regarding outcomes for each individual client. The first is a standardized evaluation sheet with intake and discharge scores in comparison with sample means and standard deviations. This also includes a graphical display of clients' intake and discharge scores as well as scores from measures of therapeutic alliance and client satisfaction. The second level is for benchmarking and displays comparisons of scores from a specific client, site, or client sample. The third level provides a graphical display of a client's trajectory of scores and includes guiding lines that indicate the baseline and action limit for the client.

Implementation of the Stuttgart-Heidelberg model fostered a clinical atmosphere of good communication regarding outcome and friendliness toward evaluation and problem solving. Another strength of the model is that in addition to alerting clinicians to clients whose scores crossed the action limit, it also alerted clinicians to clients whose assessments suggested risk for suicide (Kordy et al., 2001). The model appears effective, but the report on four years of implementing the model did not address the model's accuracy in predicting negative outcome and whether feedback to therapists improved client outcomes. Other logistical details were unclear. For example, how does the model calculate expected outcomes and how does it determine the action limit that serves as the cutoff for at-risk status?

Service profiling and outcome benchmarking. Barkham et al. (2001; cf. Mellor-Clark, Barkham, Connell, & Evans, 1999) expressed preference for "quality improvement" over "quality assurance," the latter of which may merely maintain the status quo in psychotherapeutic services. They proposed the term "quality evaluation," considering that improvement of services

depends on evaluation of existing services in comparison with valid standards. They created these standards by profiling subgroups of service settings (e.g., “secondary care settings”), of providers, and of clients (e.g., male vs. female, short vs. medium and long treatment episode durations) for their observed outcomes. The resulting profiles provided percentile benchmarks for evaluation of treatment outcome for current service locations, providers, and clients.

For outcome measurement, Barkham et al.’s (2001) system of service profiling and outcome benchmarking used periodic administration of the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE–OM), a 34-item measure assessing four domains: Subjective Well-being, Problems, Functioning, and Risk. Final outcomes on the CORE-OM can be classified as reliably changed for better or worse using a reliable change index (RCI; Jacbosen & Truax, 1991). The system can also incorporate other measures relevant to treatment concerns and context.

The service profiling and outcome benchmarking system classifies ongoing outcome in three categories. “Below clinical cutoff” indicates that the client’s score falls below the clinical cutoff score (i.e., point of division between clinical and normal range of scores; Jacobson & Truax, 1991). “Moderate” indicates that the client’s score falls above the clinical cutoff score, but below the highest quartile. “Severe” indicates that the client’s score falls within the highest quartile, which also happens to be any score higher than one standard deviation above the clinical mean. The Moderate and Severe categories presumably identify clients at risk for negative outcome, but Barkham et al. (2001) did not present prediction accuracies for an early warning function.

Fully Developed Warning Systems

The several warning systems described below appear to be fully developed in that information is readily accessible for their outcome measures, their criteria for ongoing outcomes that identify clients as at-risk, and their method of generating comparative expected outcomes, if any. The descriptions below provide information on each of these features.

Patient profiling and expected treatment response. Howard et al. (1996) and Lueger et al. (2001) are primary advocates of patient-focused research and presented an outcome monitoring system based on patient profiling and expected treatment response. This system has a fully developed early warning system component. For its outcome monitoring, the system uses periodic administration of the Mental Health Index (MHI; Howard, Brill, Lueger, O'Mahoney, & Grissom, 1995; Howard, Orlinsky, & Lueger, 1995; Sperry, Brill, Howard, & Grissom, 1996). The MHI may be completed by the client or the clinician in a computerized or paper and pencil format. Additional measures specific to treatment concerns or context may be incorporated on occasion (e.g., Presenting Problems Scale, Global Assessment Scale). The system uses a clinical cutoff score to classify final outcomes as falling in the clinical or normal range (Jacobson & Truax, 1991). It also classifies final outcomes as reliably changed for better or worse using a reliable improvement index (RII, a variant on RCI; cf. Jacobson & Truax, 1991).

The system creates profiles of ongoing MHI scores for individual clients and identifies clients at risk for negative outcome when scores deviate from their expected treatment response. Deviation reaches an at-risk magnitude when scores cross a rationally derived 25% failure boundary. This boundary is one side of a confidence interval around the expected scores for any given client and indicates that only 25% of clients with similar characteristics would have a score deviating to such an extreme at that particular time in treatment. Thus the system identifies at-

risk clients by comparing actual outcomes with expected outcomes. These expected outcomes are specific to each client because they are generated using client-specific variables as part of hierarchical linear models (HLM; Bryk & Raudenbush, 1992). Howard et al. (1996) originally identified 18 such client variables, which Lutz, Martinovich, and Howard (1999) later narrowed down to seven: current well-being, current symptoms, current life functioning, clinician-rated severity, chronicity, previous treatment, and treatment expectation. These pre-treatment predictors accounted for 22% of variability in rates of change.

Howard et al. (1996) based their view of expected treatment response on dosage and phase models for how much symptoms improve (i.e., response) per session of treatment (i.e., dose; Howard et al., 1986). They observed a curvilinear change trajectory, with treatment responses that were large initially and smaller later on (i.e., a curvilinear change trajectory that begins steep and levels off over time in treatment). They attributed the curvilinearity—varying rates of response—to three sequential phases that clients pass through during treatment (Howard, Lueger, Maling, & Martinovich, 1993).

In the Remoralization phase, clients entering therapy may be particularly demoralized (Frank & Frank, 1991) by their problems yet may respond quickly to therapy. This corresponds to the steep initial part of the change trajectory and typically lasts only several sessions. In the Remediation phase, interventions attempt to remediate symptoms and shift the client toward coping skills that are more effective in relieving symptoms. This corresponds to a moderately steep central portion of the change trajectory and lasts approximately 16 sessions (Kopta, Howard, Lowry, & Beutler, 1994). The final phase, Rehabilitation, reflects more typical psychotherapy, a gradual and deeper-level process of replacing maladaptive behaviors with those that are adaptive. This corresponds to a nearly flat latter portion of the change trajectory and has

a duration dependent upon the severity and nature of the treatment concern (Maling, Gurtman, & Howard, 1995). The MHI has subscales tapping into phenomena specific to each of these three phases: subjective well-being, symptoms, and life functioning, respectively.

The warning system based on patient profiling and expected treatment response provided therapists with treatment progress reports for each client. These reports included three to four pages (computerized or printed) of text or graphics summarizing client characteristics, presenting problems, MHI tracking data, progress on MHI components, MHI percentile ranking as a function of sessions, and current overall change score (i.e., difference between current score and baseline). Graphical displays included overlaid plots of clients' ongoing outcome, clients' expected outcome, clinical cutoff scores, and the 25% failure boundary. This warning system identified 88% of actual deteriorators and appropriately classified 82% of non-deteriorators using a criterion of non-improvement on the current symptoms subscale by session 12 (according to the RII). Using a criterion of two consecutive scores exceeding the 25% failure boundary, the system identified 64–76% of deteriorators.

A strength of this system is that its expected treatment response models have theoretical basis in dosage response and phase models. Another interesting strength is how the system predicts various likelihoods of particular final outcomes given certain midtreatment outcomes. For example, the system indicates that clients who fail to remoralize after four sessions have a 50% likelihood of treatment failure. However, the multiple and varied criteria for outcome predictions may cause the system to be somewhat unwieldy for therapists. The system's computerization may handle these complexities automatically, but also may make the system less accessible to providers for whom incorporation of specialized software is inconvenient. It may also be a concern that many of the system's predictions take place—or are at their highest

accuracy—after 12 treatment sessions. This may be too late to identify at-risk clients before they drop out of treatment and may leave too little time to influence their trajectory.

There appears to be no report of whether this system's feedback to therapists yields improved outcomes for clients. In addition, the inconsistent and periodic administration of the outcome measures may produce more compliance challenges. Routine session-by-session administration could improve compliance and could also yield a more accurate and detailed profile of client outcome.

In terms of prediction accuracy, the reported 22% of variability in trajectory slopes accounted for by the model's seven predictor variables may be confounded. The potential problem is that three of the predictor variables are intake scores on the MHI's three subscales, but they also combine to be the MHI total score, which is the variable being predicted. In other words, these three independent variables are the same as one data point from the dependent variable on the other side of the model's equation (i.e., intake MHI score). It appears to be a client of some data predicting themselves, which could inflate estimates of variability accounted for by the model's predictors.

OQ system. The OQ system for outcome monitoring and early warning (Finch et al., 2001; Lambert, Hansen, et al., 2001) stems from the outcome research of Michael Lambert and Gary Burlingame (see www.oqmeasures.com). To monitor outcomes, the system uses session-by-session administration of the Outcome Questionnaire (OQ-45; Lambert et al., 2004). The OQ-45 is a 45-item self-report measure available in computerized or paper and pencil format. It has demonstrated high reliability, validity, and sensitivity to change.

The OQ system is a product of research regarding expected outcomes for clients in psychotherapy (Anderson & Lambert, 2001; Hansen, 1999; Kadera, Lambert, & Andrews,

1996). It monitors outcomes for purposes of treatment planning and quality care. It informs clinicians regarding client progress of any type (e.g., improvement or deterioration) and also identifies clients at risk for negative outcome. The system uses a clinical cutoff score to indicate whether scores fall in the clinical or normal range. It also uses a reliable change index (RCI; Jacobson & Truax, 1991) to identify final scores that are reliably changed for better or worse. The system's feedback to clinicians is immediate so that they can make inquiries or adjustments based on clients' current scores. The feedback may be computerized or on printed pages and typically involves a textual feedback message and graphical display of plotted actual scores, expected scores, and the clinical cutoff score.

The early warning system has used two different methods of identifying clients at risk for negative outcome (Lambert et al., 2002). The original method was developed by expert judges and involves rationally derived algorithms for the amount of negative deviation that must occur by a given session. The second method of identifying at-risk clients involves empirically derived algorithms. The empirical approach compares actual scores to expected scores as modeled by hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). Given that expected change trajectories vary by initial OQ-45 scores, the OQ system uses comparison trajectories created using data from clients with similar intake scores. Clients are signaled as at-risk if their scores exceed a threshold indicating that their deviation is within the most extreme 10% of deviating clients, this percentage corresponding to the deterioration rate in adult clients. This threshold is the boundary on one side of a confidence interval created around the expected change trajectory scores.

In one study, the OQ system's accuracy in predicting which clients would have negative outcomes was somewhat higher using the empirically versus the rationally derived algorithms

(Lambert et al., 2002; Spielmans, Masters, & Lambert, 2006). The system's hit rate for distinguishing deteriorators from non-deteriorators was 79–83%. The rational method's sensitivity in identifying actual deteriorators was 81%, whereas the empirical method had a sensitivity of 83% by the third session and 100% overall. The system's predictions included 17–21% of clients as false positives for deterioration, but this may not be a problem considering that most of these clients were non-responders and would likely have benefited from extra clinical attention.

A strength of the OQ system is that it encourages the administration of additional measures when it alerts therapists to at-risk clients. These Clinical Support Tools provide the therapist additional insight into the clients' situation (e.g., therapeutic alliance, client motivation to change, client social support network, client perfectionism, and client stressful life events). Clients whose therapists received feedback from the OQ system have experienced improved outcomes (Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert, Whipple, et al., 2001; Lambert et al., 2002; Whipple et al., 2003). Compared to at-risk clients in the nonfeedback condition, nearly twice as many at-risk clients from the feedback condition ended treatment with improvement (9 clients vs. 4) and even more ended with recovery (i.e., final scores in the nonclinical range; 5 clients vs. 1). These superior outcomes may be due to the at-risk clients in the feedback condition receiving twice as many sessions on average (9.3 sessions vs. 4.7), presumably as a result of the feedback. In addition, it appears that simultaneous feedback to therapists and their clients may achieve even better outcomes than when only therapists receive feedback (Hawkins et al., 2004).

Another strength of the OQ system is that its feedback is immediate. It is standard for clients to complete the OQ-45 upon presenting at a treatment session and the therapist to have

the scores and feedback as clients enter the therapy room. In addition, the warning system is accurate even in the early stages of treatment, which is crucial in identifying at-risk clients before they drop out or get too far along on a path of deterioration. Although the OQ system has software available, it need not be computerized. In a noncomputerized approach, a therapist could photocopy a graph for the appropriate expected change trajectory (based on initial score), put it in the client's chart, and then plot the client's OQ-45 scores throughout treatment, attending to whether scores exceed the 10% threshold for at-risk clients. Expected trajectories do not need to be recalculated for each client because they are merely mean trajectories based on initial scores. This simplicity increases the likelihood that clinicians can easily use the system in routine practice (Lambert et al., 2002). This form of outcome monitoring could help focus case managers' attention to the roughly 10% of clients at risk for negative outcome and relieve them from such close attention to other clients (Finch et al., 2001).

The success of the OQ system's model has been replicated with the Youth Outcome Questionnaire-30 and the Youth Outcome Questionnaire-64, both of which are youth versions of the OQ-45 (Bybee et al., 2007; Cannon et al., 2010). The OQ system's feasibility for routine clinical practice has been demonstrated as well (Lambert, Hansen, et al., 2001). Although the simplicity of using a single outcome measure affords the OQ system its feasibility, a single measure may not assess all relevant aspects of treatment for all clients. In addition, repeated administration of self-report outcome measures may result in unreliable responding habits.

Early Warning Systems and Managed Care

The managed care industry has taken interest in systems of outcome monitoring to inform practice guidelines, client satisfaction, and efforts in cost-effectiveness (Mordock, 2000; Sharfstein & Stoline, 2000). Such interests and efforts are not limited to the United States

(Barkham et al., 2001; Kordy et al., 2001). In terms of cost-containment, third-party payors are particularly interested in better understanding treatment effectiveness across time (Bloom, 1987; Brokowsky, 1991; Richardson & Austad, 1991; Sabin, 1991). Some third-party payors base their authorizations of treatment type and amount on data from outcome measures (Mirin & Namerow, 1991; Moses-Zirkes, 1994). This customization of authorizations achieves cost-efficiency as well as flexibility based on symptom levels, symptom types, setting of care (e.g., managed care vs. community mental health system; Warren et al., 2010), and other client variables associated with change.

Managed care organizations face the criticism of providing treatment at only minimum levels in order to cut costs (Docherty, 1999; Miller, 1996). In response, these organizations are increasingly using patient-focused outcome monitoring to ensure quality while minimizing costs (O'Donahue, Graczyk, & Yeater, 1998). Identification of at-risk clients using outcome monitoring typically increases quality of care for these clients and helps them receive appropriate services. Outcome monitoring could also serve to identify providers who achieve superior outcomes for their clients (Matsumoto, Jones & Brown, 2003). This identification could increase therapist productivity, acting as an alternative or an addition to incentive programs that are the more typical tool used to boost productivity (Bobbitt, Marques, & Trout, 1998; Gunn, 1998). However, outcome measures are more commonly used for in-house studies of treatment effectiveness rather than for identifying effective providers (Steenbarger & Smith, 1996) and there may be confounds to the latter usage (e.g., therapists may achieve differing outcomes due to systematic differences in clientele rather than due to personal capacity for productivity).

Johnson and Shaha (1996) contrasted quality assurance with Continuous Quality Improvement in managed care. Quality assurance is primarily an external evaluation imposed on

providers and may focus more on what is easily quantified, such as provider qualifications (e.g., degree and licensure, documentation of adherence to protocol, and number of malpractice claims) as opposed to quality of care. Quality assurance ensures qualification and procedure, which may indirectly ensure a certain level or quality of care, but may primarily guarantee administrative and procedural burden.

Continuous Quality Improvement, in contrast, involves internal evaluation of quality using methods developed from within the clinical setting. This approach has a greater likelihood of improving quality of care. Outcome measures that are sensitive to change could play an integral role in Continuous Quality Improvement, as could measures of customer satisfaction and therapeutic relationship (Johnson & Shaha, 1996). Outcome monitoring systems and early warning systems are good examples of Continuous Quality Improvement and have improved client outcomes (Harmon et al., 2007; Hawkins et al., 2004; Lambert, Hansen, et al., 2001; Lambert et al., 2002; Whipple et al., 2003). As mentioned above, one system experimented with simultaneous feedback to clinicians and clients and achieved greater symptom reduction than when only the clinicians received the feedback (Hawkins et al., 2004). These are examples of studies in the realm of managed care and evidence-based practice that have recently begun to examine individuals' negative responses to psychotherapy as opposed to examining treatments whose effects appear negative (Lilienfeld, 2007).

Outcome Research and Early Warning Systems for Youth

As described above, research literature for adult psychotherapy features exciting advances in outcome tracking and early identification of clients at risk for negative outcome. These advances improve outcomes for all clients and especially help clinicians and managed care organizations prevent treatment non-responders from experiencing negative outcome. The

literature for children and adolescents has lagged behind adult research (Durlak & McGlinchey, 1999; Kazdin, 2003). The scarcity of outcome monitoring and early warning systems for youth is particularly unfortunate because youth deterioration rates may be higher than rates for adults (Bishop et al., 2005; Cannon et al., 2010; Weisz, Donenberg, et al., 1995). In addition, effect sizes are near zero for youth treatments in some settings (Weisz, Donenberg, et al., 1995) and 40–60% of youth drop out of treatment early (Kazdin, 2003; Wierzbicki & Pekarik, 1993).

Nonetheless, the outlook is good for youth research and practice because outcome research is broadening and growing (Durlak & McGlinchey, 1999; Kazdin, 2003) and therapy appears beneficial in general (Casey & Berman, 1985; Kazdin, Bass, Ayers, & Rodgers, 1990; Weisz, Weiss, Han, Granger, & Morton, 1995). Although some effect sizes are poor, general effect sizes for youth approximate those of adult populations (Durlak & McGlinchey, 1999; Weisz, Weiss, & Donenberg, 1992) and individual and group therapies for youth are comparable in effectiveness (Hoag & Burlingame, 1997). However, given the generalizability problems of efficacy and effectiveness research described above, it may be appropriate to temper estimates of effectiveness (Weisz et al., 1992). Similarly, Kazdin (2003) points out that clients in typical clinical settings may have lower distress levels than in the clinical trials, further compromising generalizability.

Regarding the outcome monitoring and early identification of at-risk clients, Kazdin (2005) noted that “such information would be enormously helpful if used to monitor and evaluate treatment in clinical practice” (p. 555). Early warning systems for youth would be particularly helpful considering estimated premature dropout rates of 40–60% (Kazdin, 2003; Wierzbicki & Pekarik, 1993). Pekarik and Stephenson (1988) found adult dropout to be related to therapist experience and referral source, but their study found no predictive variables for youth

dropout. They found that youth dropout occurred after nearly twice as many treatment sessions as adult dropout, a delay likely attributable to the termination decision not falling on the primary client, as with adult treatment. Instead, the decision to terminate falls on these youths' parents, who may be slightly removed from the therapy process. One study identified parent self-criticism and delusional guilt to be a predictor of child dropout (Venable & Thompson, 1998).

There have been several studies testing the accuracy of early warning systems for identifying youth at risk for negative treatment outcome. These studies are based on the OQ system described above. Bishop et al. (2005) reported a study monitoring outcomes using the Youth-Outcome Questionnaire-64 (YOQ-64; Burlingame et al., 2005), a youth version of the Outcome Questionnaire (OQ-45; Lambert et al., 2004). The study sample included 300 youth ages 3–18. To identify clients at risk for negative outcome, this early warning system used rationally derived algorithms for the amount of negative deviation that must occur by a given session. The warning system identified 77% of the deteriorators overall, with higher sensitivity for predicting deteriorators in the residential setting.

Bybee et al. (2007) reported a study testing the prediction accuracy of a similar outcome monitoring and early warning system. This study tracked outcome using periodic administration of the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004), a shortened version of the 64-item Youth Outcome Questionnaire (YOQ-64; Burlingame et al., 2005). This system used empirically derived algorithms to identify clients at risk for negative outcome, in a similar manner to the OQ system described above. The empirical approach compares actual scores to expected scores as modeled by hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002). Given that expected change trajectories vary by initial YOQ scores, this system uses comparison trajectories created using data from clients with similar intake scores. Clients are signaled as at-

risk if their scores exceed a threshold indicating that their deviation is within the most extreme 10% of deviating clients, this percentage corresponding to the researchers' estimated deterioration rate for youth clients. This threshold is the boundary on one side of a confidence interval created around the expected change trajectory scores. The warning system identified 72% of the deteriorators. A potential limitation to the study was that it did not control for its usage of both self-report and parent-report YOQ scores, which may show some systematic differences. In addition, the expected trajectories did not control for covariates other than initial score.

Cannon et al. (2010) tested for systematic differences in self-report versus parent-report scores on the YOQ-64 by examining hierarchical linear models for each, and controlling for the effects of covariates. Self-report change trajectories had a slightly lower elevation and faster rate of change than parent-report trajectories. This study's warning system used the YOQ-64 as its outcome measure and used empirically derived algorithms for identifying at-risk clients (cf. Bybee et al., 2007). The system's accuracy using self-report YOQ-64 scores to predict clients with negative outcome was comparable to its accuracy using parent-report scores. The system's accuracy was highest when it simultaneously used self-report and parent-report YOQ-64s, identifying 70% of deteriorators.

Warren et al. (2010) also examined YOQ-64 scores, but tested for difference in trajectories for clients treated in a community mental health system versus a large managed care setting. They demonstrated that the managed care setting had lower initial symptom severity and faster rates of improvement. Similar to Cannon et al. (2010) and Bybee et al. (2007), the warning system of this study used the empirically derived algorithms for identifying at-risk clients. The warning system identified 84% of deteriorators in the community system but only 58% in the

managed care setting. Clients signaled as at-risk were 7.3 or 3.4 times more likely to end in deterioration than not (in the community and managed care settings, respectively).

As demonstrated by the aforementioned studies of youth outcome monitoring and early warning systems, the youth research literature is making great progress toward improving outcomes for youth in psychotherapy treatment. Important youth research has yet to be accomplished, however. For example, future studies could replicate the above prediction accuracies, perhaps using differing measures or populations. Future studies could also replicate or find alternatives to the variables predictive of youth change trajectories. The predictive variables from the Bybee et al. (2007), Cannon et al. (2010), and Warren et al. (2010) studies included initial score, prior psychotherapy treatment, age, total number of weeks in treatment, self-report versus parent-report, and community mental health setting versus managed care setting (the variables were not all used simultaneously). Ultimately, future studies will test whether implementation of the warning system with feedback to therapists improves outcomes for youth clients.

Present Study

To review, the field of mental health treatment is making efforts to better serve all psychotherapy clients, but especially the 5–10% of clients who deteriorate in treatment (Lambert & Ogles, 2004) and the 30–60% who drop out prematurely (Pekarik & Stephenson, 1988). These efforts involve collaboration between research and practice because therapists on their own are less accurate in predicting which clients will experience negative outcome. This collaboration between research and practice has required bridging the divide that has existed between the two. Both research and practice have been treatment focused for much of their history, primarily examining treatment efficacy or effectiveness, and never quite settling on the generalizability or

applicability of specific treatments. The patient-focused research paradigm has shifted the focus from treatment outcomes on the group level to outcomes on the individual client level. This movement involves outcome monitoring for purposes of treatment planning and quality care. Some of these monitoring systems include early warning systems that could help identify and better serve clients who are at risk for negative outcome.

The present study attempted to take an important next step in the development of outcome monitoring and early warning systems for youth by validating previous studies and replicating tests for variables that were predictive of youth change trajectories. This study also replicated the accuracy of a warning system for at-risk youth clients, using the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004). The results from this study contribute to the understanding and application of warning systems to clinical settings for youth. In response, future studies could compare outcomes between client groups whose therapists do or do not receive systematic feedback. This endeavor offers many benefits to quality improvement efforts being made by clinicians and managed care organizations.

This study intended to contribute to the psychotherapy research literature that is developing outcome monitoring and early warning systems to better serve youth clients. The first aim was to develop change trajectories for the YOQ scores over time, identifying any variables predictive of expected change trajectories. These trajectories inform the research literature as to what patterns of change may be expected and which variables seem to have an impact on these patterns. Similar trajectory models played an integral role in accomplishing the second aim of this study, which was to calculate the accuracy of a warning system identifying clients at risk for negative outcome. Similar to past studies described above, these predictions were based on how the scores compare to prediction intervals around expected trajectories.

METHOD

This study examined archival data for a brief psychotherapy outcome measure administered to youth in a large private managed care organization. In the first part of the study, we identified client variables associated with outcome scores over time. We also calculated the variability in outcome scores associated with differences in clients, therapists, and treatment sites. In the second part of the study, we created cutoffs to identify which ongoing outcome scores reached a severity predictive of negative final outcomes. We then tested the accuracy of the resulting predictions in order to demonstrate the accuracy an early warning system could potentially attain if implemented in clinical practice to identify youth at risk for negative outcome.

Participants and Procedure

This study analyzed data selected from the archives (1999–2005) of a large private managed care organization providing services throughout the United States. Clients seeking outpatient psychotherapy services through this organization were typically of average to above-average socioeconomic status. The organization's mental health providers included psychiatrists, psychologists, social workers, marriage and family therapists, and others. Mental health services for youth primarily included individual and family psychotherapy and medication management visits. Clinicians used various therapeutic approaches in these visits, with family therapy and cognitive strategies being common with youth clients. Data were collected as part of routine services at the first, third, and fifth sessions, and then once every five sessions or fewer. Youth or their parents or guardians completed the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004) at check-in when presenting for outpatient treatment, typically requiring 5 minutes or less.

Table 1 demonstrates our steps in selecting client data from the reliable data of the original archive. We began with data for 15,460 clients having valid values for sex and age and for whom the treatment episodes were confirmed as complete, based on our decision to let 90+ days of no contact mark the end of a treatment episode. In our second step, we selected data for clients with a YOQ measurement within the first two sessions of treatment. Only the service types with a psychotherapy component were counted as sessions of psychotherapy treatment. Table 2 identifies these specific services by their current procedural terminology codes. In our third step, we selected data for clients who had a YOQ near the end of treatment (no more than three sessions or seven weeks of treatment after final YOQ). In our fourth step, we selected clients with at least two YOQ measurements and at least 2 sessions of treatment. With a final step of selecting data for clients with episode lengths that did not exceed the 90th percentile (26 sessions), we arrived at our sample of 4,309 clients for the analyses of part1 of the study, comprising 38% of the original reliable data in the archive.

Table 3 presents the demographics for the sample selected for the analyses of Part 1 of this study. This sample of 4,309 clients was 37% female, with a mean age of 9.4 years old. Table 4 shows that adjustment disorders were the most common primary diagnosis for this sample (35%), followed by attention-deficit/hyperactivity disorders (19%) and mood disorders (15%). At least 8% of clients had multiple diagnoses on record. Table 3 shows that there were 1,637 therapists on record for these clients, apparently primarily psychologists (18%), marriage and family therapists (16%), social workers (11%), and medical doctors (5%). The degrees or credentials for the other therapists were unknown (50%).

We used *t* tests (see Table 5) and chi-square tests (see Table 6) to identify significant differences between this selected sample and the original archive. Most variables were

Table 1

Steps Taken in Sample Selection Process

Step	*N*	Percent of archive	Selection criteria
Step 1	15,460	100%	Valid values for sex and age. Treatment complete (no treatment sessions for 90 days).
Step 2	11,160	72%	1st YOQ within first 2 sessions.
Step 3	5,733	37%	No more than 3 sessions or 7 weeks in treatment after last YOQ.
Step 4	4,542	29%	At least 2 YOQs and 2 sessions of treatment.
Sample Part 1	4,309	38%	No episodes longer than 26 sessions (90th percentile).
Sample Part 2	1,744	11%	At least 3 YOQs and 3 sessions of treatment.

Table 2

Current Procedural Terminology (CPT) Codes Qualifying as Psychotherapy

CPT Code	Description
Psychotherapy treatment	
90804	Individual psychotherapy, office, 20–30 min
90806	Individual psychotherapy, office, 45–50 min
90808	Individual psychotherapy, office, 75–80 min
90810	Individual psychotherapy, office, interactive, 20–30 min
90812	Individual psychotherapy, office, interactive, 45–50 min
90814	Individual psychotherapy, interactive, office, 75–80 min
90843	Outdated code replaced by 90804
90844	Outdated code replaced by 90806
Psychotherapy with medication management	
90805	Individual psychotherapy, office, 20–30 min; w/E&M
90807	Individual psychotherapy, office, 45–50 min; w/E&M
90809	Individual psychotherapy, office, 75–80 min; w/E&M
90811	Individual psychotherapy, office, interactive, 20–30 min; w/E&M
90813	Individual psychotherapy, office, interactive, 45–50 min; w/E&M
90815	Individual psychotherapy, office, interactive, 75–80 min; w/E&M
Other	
90845	Psychoanalysis
90847	Family psychotherapy (conjoint psychotherapy) (w/patient present)

Table 3

Descriptive Statistics for Part 1 Sample

Characteristic	*M*	*SD*	*Mdn*	Range	Characteristic	*n*	%
n YOQs per client	2.7	1.2	2.0	2–15	Female	1,568	36.4%
Weeks between YOQs	6.3	4.6	5.1	1–62	From day tx/ inpatient	62	1.4%
Sessions between YOQs	2.8	1.5	2.5	0–13	Prior treatment	658	15.3%
Treatment episode number	1.2	0.7	1.0	1–10	Straight from inpatient	35	0.8%
Treatment episode length (weeks)	17.4	15.5	13.0	1–172	Straight from day tx	27	0.6%
Treatment episode length (sessions)	7.6	5.0	6.0	2–26	Fully nested w/i site	4,241	98.4%
Age	9.4	2.7	9.2	4–17	Fully nested w/i ther	3,818	88.6%
Change score	-3.5	13.9	-3.0	-76–101	Therapist sex		
Sessions before 1st YOQ	1.0	0.8	1.0	0–3	Female	560	34.2%
Baseline YOQ	41.1	17.5	40.0	0–109	Male	352	21.5%
Sessions per month	2.5	1.4	2.2	0–14	Data missing	725	44.3%
YOQs per month	1.1	0.8	0.9	0–9	Therapist degree		
Therapist year of practice (*n* = 550; *n* missing = 1,087)	22.6	8.3	22.6	4–52	PhD	298	18.2%
Therapist age (*n* = 754; *n* missing = 883)	54.1	7.8	53.7	31–79	MFT	258	15.8%
					SW	184	11.2%
					MD	78	4.8%
					Other/unknown	819	50.0%

Note. *N* = 4,309. PhD = psychologists. MFT = marriage and family therapists. SW = social workers. MD = medical doctors.

Table 4

Primary Diagnoses for Part 1 Sample

Primary diagnoses	*n*	%	Primary diagnoses	*n*	%
Adjustment disorders	1,518	35.2%	Conduct disorders	151	3.5%
Attention-deficit/hyperactivity disorders	835	19.4%	Posttraumatic stress disorder	104	2.4%
Mood disorders	645	15.0%	Abuse/neglect of child	11	0.3%
Anxiety-related disorders	440	10.2%	Autistic disorders	70	1.6%
Oppositional defiant disorder	280	6.5%	Substance abuse/dependence	6	0.1%
Other/unknown	249	5.8%			

Note. $N = 4{,}309$. Eight percent of clients had multiple diagnoses appearing in their insurance claims data. Comorbidity rates may have been higher.

Table 5

Comparing Part 1 Sample to Archive: t *Tests*

	Selected sample[a]		Archive[b]		Sample comparisons		
Characteristic	*M*	*SD*	*M*	*SD*	*t*	*df*	Method
Baseline YOQ	41.1	17.5	43.0	17.8	-6.07*	19,767	pooled
Episode number	1.2	0.7	1.3	0.8	-2.53*	7,503	Satterthwaite
Treatment episode length (sessions)	7.6	5.0	10.9	7.3	-34.62*	10,016	Satterthwaite
Treatment episode length (weeks)	17.4	15.5	24.8	20.6	-25.57*	9,006	Satterthwaite
Age	9.4	2.7	10.5	3.2	-22.18*	7,957	Satterthwaite
n YOQs per client	2.7	1.2	2.1	1.3	31.55*	7,457	Satterthwaite
Sessions before 1st YOQ	1.0	0.8	2.8	3.8	-52.58*	19,037	Satterthwaite
Weeks between YOQs	6.3	4.6	14.6	14.4	-61.12*	19,508	Satterthwaite
Sessions between YOQs	2.8	1.5	6.4	5.4	-74.06*	19,762	Satterthwaite
Change score	-3.5	13.9	-1.8	10.8	-7.41*	5,834	Satterthwaite
Sessions per month	2.5	1.4	2.4	1.5	.2.35*	7,657	Satterthwaite
YOQs per month	1.1	0.8	0.6	0.9	29.69*	19,730	pooled

[a]n = 4,309. [b]n = 15,460.

*$p < .05$.

Table 6

Comparing Part 1 Sample to Archive: Chi-Square Tests

	Selected sample[a]		Archive[b]		Sample comparisons	
Characteristic	*n*	%	*N*	%	χ^2	*df*
Female	1,568	36.4%	6,073	39.3%	11.89*	1
From day tx/ inpatient	62	1.4%	396	2.6%	18.77*	1
Prior treatment	658	15.3%	2,598	16.8%	5.77*	1
Straight from inpatient	35	0.8%	283	1.8%	22.08*	1
Straight from day tx	27	0.6%	113	0.7%	0.52	1
Fully nested w/i site	4,241	98.4%	15,116	97.8%	6.91*	1
Fully nested w/i ther	3,818	88.6%	12,669	82.0%	107.90*	1

[a]n = 4,309. [b]n = 15,460.

*$p < .05$.

significantly different between the two samples, likely due to the high statistical power available in detecting differences with such large sample sizes. The more notable differences between the samples were expected given our selection criteria (e.g., selected sample with shorter treatment episodes, more YOQs per client, and first YOQ earlier in treatment). No differences appeared too dramatic.

In an additional step of selecting data for clients with at least 3 YOQ measurements and at least 3 sessions of treatment, we arrived at our sample of 1,744 clients for the analyses of Part 2 of the study, comprising 11% or the original reliable data in the archive. Table 7 presents the demographics for this second sample and Table 8 presents the primary diagnoses. The sample characteristics were fairly similar to those of Part 1, just with a smaller sample size of 1,744 clients. We used *t* tests (see Table 9) and chi-square tests (see Table 10) to identify significant differences between this selected sample for Part 2 of the study and the original archive. Most variables were different between the two, likely due to the high statistical power available in detecting differences with such large sample sizes. The more notable differences between the samples were expected given our selection criteria (e.g., more frequent YOQ administration). No differences appeared too dramatic.

We also compared the Part 1 sample with the smaller Part 2 sample. The selection criteria that distinguished the two samples were that the Part 1 clients had two or more sessions and two or more YOQs whereas Part 2 clients had three or more of each. Table 11 presents the results for the related *t* tests and Table 12 presents the results of the related chi-square tests. Given these different criteria, the expected sample differences were that clients in the Part 2 sample had longer treatment episodes (in terms of sessions and weeks), more YOQs per client, and larger overall change scores for the YOQ. Less obvious, yet still sensible, is that the Part 1 sample had

Table 7

Descriptive Statistics for Part 2 Sample

Characteristic	*M*	*SD*	*Mdn*	Range	Characteristic	*n*	%
n YOQs per client	3.7	1.2	3.0	3–15	Female	598	34.3%
Weeks between YOQs	6.7	4.0	5.9	1–30	From day tx/ inpatient	31	1.8%
Sessions between YOQs	2.9	1.3	2.7	0–8	Prior treatment	300	17.2%
Treatment episode number	1.3	0.8	1.0	1–10	Straight from inpatient	19	1.1%
Treatment episode length (weeks)	25.4	18.4	21.0	2–172	Straight from day tx	12	0.7%
Treatment episode length (sessions)	10.8	5.5	10.0	3–26	Fully nested w/i site	1,712	98.2%
Age	9.2	2.6	9.1	4–17	Fully nested w/i ther	1,487	85.3%
Change score	-4.5	15.3	-4.0	-76–101	Therapist sex		
Sessions before 1st YOQ	1.0	0.8	1.0	0–3	Female	333	35.7%
Baseline YOQ	42.1	17.5	41.0	0–104	Male	199	21.3%
Sessions per month	2.3	1.1	2.1	0–11	Data missing	402	43.0%
YOQs per month	0.9	0.6	0.7	0–7	Therapist degree		
Therapist year of practice (*n* = 316; *n* missing = 618)	22.6	8.6	22.6	4–52	PhD	174	18.6%
Therapist age (*n* = 424; *n* missing = 507)	54.1	7.7	53.8	34–79	MFT	153	16.4%
					SW	97	10.4%
					MD	47	5.0%
					Other/unknown	463	49.6%

Note. *N* = 1,744. PhD = psychologists. MFT = marriage and family therapists. SW = social workers. MD = medical doctors.

Table 8

Primary Diagnoses for Part 2 Sample

Primary diagnoses	*n*	%	Primary diagnoses	*n*	%
Adjustment disorders	566	32.5%	Conduct disorders	49	2.8%
Attention-deficit/hyperactivity disorders	349	20.0%	Posttraumatic stress disorder	51	2.9%
Mood disorders	315	18.1%	Abuse/neglect of child	4	0.2%
Anxiety-related disorders	175	10.0%	Autistic disorders	29	1.7%
Oppositional defiant disorder	131	7.5%	Substance abuse/dependence	0	0.0%
Other/unknown	75	4.3%			

Note. N = 1,744. Twelve percent of clients had multiple diagnoses appearing in their insurance claims data. Comorbidity rates may have been higher.

Table 9

Comparing Part 2 Sample to Archive: t *Tests*

	Selected sample[a]		Archive[b]		Sample comparisons		
Characteristic	*M*	*SD*	*M*	*SD*	*t*	*df*	Method
Baseline YOQ	42.1	17.5	43.0	17.8	-2.02*	17,202	pooled
Episode number	1.3	0.8	1.3	0.8	-0.00	17,202	pooled
Treatment episode length (sessions)	10.8	5.5	10.9	7.3	-0.63	2,489	Satterthwaite
Treatment episode length (weeks)	25.4	18.4	24.8	20.6	1.36	2,265	Satterthwaite
Age	9.2	2.6	10.5	3.2	-18.41*	2,378	Satterthwaite
n YOQs per client	3.7	1.2	2.1	1.3	52.43*	17,202	pooled
Sessions before 1st YOQ	1.0	0.8	2.8	3.8	-48.76*	12,650	Satterthwaite
Weeks between YOQs	6.7	4.0	14.6	14.4	-52.19*	8,587	Satterthwaite
Sessions between YOQs	2.9	1.3	6.4	5.4	-65.04*	10,266	Satterthwaite
Change score	-4.5	15.3	-1.8	10.8	-7.36*	1,946	Satterthwaite
Sessions per month	2.3	1.1	2.4	1.5	-5.22*	2,596	Satterthwaite
YOQs per month	0.9	0.6	0.6	0.9	16.48*	2,658	Satterthwaite

[a]n = 1,744. [b]n = 15,460.

*$p < .05$.

Table 10

Comparing Part 2 Sample to Archive: Chi-Square Tests

	Selected sample[a]		Archive[b]		Sample comparisons	
Characteristic	*n*	%	*n*	%	χ^2	*df*
Female	598	34.3%	6,073	39.3%	16.46*	1
From day tx/ inpatient	31	1.8%	396	2.6%	4.59*	1
Prior treatment	300	17.2%	2,598	16.8%	0.00	1
Straight from inpatient	19	1.1%	283	1.8%	5.81*	1
Straight from day tx	12	0.7%	113	0.7%	0.04	1
Fully nested w/i site	1,712	98.2%	15,116	97.8%	1.12	1
Fully nested w/i ther	1,487	85.3%	12,669	82.0%	11.83*	1

[a]n = 1,744. [b]n = 15,460.

*p < .05.

Table 11

Comparing Samples for Part 1 and Part 2: t *Tests*

	Part 1 sample[a]		Part 2 sample[b]		Sample comparisons		
Characteristic	*M*	*SD*	*M*	*SD*	*t*	*df*	Method
Baseline YOQ	41.1	17.5	42.1	17.5	-1.91	6,051	pooled
Episode number	1.2	0.7	1.3	0.8	-1.48	2,997	Satterthwaite
Treatment episode length (sessions)	7.6	5.0	10.8	5.5	-21.21*	2,953	Satterthwaite
Treatment episode length (weeks)	17.4	15.5	25.4	18.4	-16.02*	2,789	Satterthwaite
Age	9.4	2.7	9.2	2.6	2.17*	3,348	Satterthwaite
n YOQs per client	2.7	1.2	3.7	1.2	-30.09*	3,046	Satterthwaite
Sessions before 1st YOQ	1.0	0.8	1.0	0.8	1.30	6,051	pooled
Weeks between YOQs	6.3	4.6	6.7	4.0	-3.67*	3,678	Satterthwaite
Sessions between YOQs	2.8	1.5	2.9	1.3	-3.09*	3,561	Satterthwaite
Change score	-3.5	13.9	-4.5	15.3	2.52*	2,980	Satterthwaite
Sessions per month	2.5	1.4	2.3	1.1	6.22*	3,982	Satterthwaite
YOQs per month	1.1	0.8	0.9	0.6	9.32*	4,578	Satterthwaite

[a]n = 4,309. [b]n = 1,744.

*p < .05.

Table 12

Comparing Samples for Part 1 and Part 2: Chi-Square Tests

	Part 1 sample[a]		Part 2 sample[b]		Sample comparisons	
Characteristic	*n*	%	*n*	%	χ^2	*df*
Female	1,568	36.4	598	34.3	2.38	1
From day tx/ inpatient	62	1.4	31	1.8	0.66	1
Prior treatment	658	15.3	300	17.2	2.04	1
Straight from inpatient	35	0.8	19	1.1	0.69	1
Straight from day tx	27	0.6	12	0.7	0.07	1
Fully nested w/i site	4,241	98.4	1,712	98.2	0.50	1
Fully nested w/i ther	3,818	88.6	1,487	85.3	12.80*	1

[a]n = 4,309. [b]n = 1,744.

*p < .05.

more YOQs per month. The Part 1 sample included clients with fewer YOQs, which on the majority should correspond to clients with fewer sessions in treatment, and it was during those early sessions that the YOQ is administered most frequently (i.e., at sessions 1, 3, 5, and at every fifth session or fewer after that). One potential explanation for the greater number of sessions per month in the Part 1 sample could follow a similar line of reasoning; early stages of treatment likely correspond to higher session frequency. The Part 1 sample included more clients in early stages of treatment (i.e., 2+ sessions, vs. the 3+ sessions of the Part 2 sample). Similarly, the Part 2 sample's higher percentage of clients not fully nested within therapists (i.e., with more than one therapist) may be expected given that longer treatment episodes offer more opportunity for a change in therapist. The other difference was in the mean age in each sample, 9.4 years in the Part 1 sample versus 9.2 in the Part 2 sample.

Measure

The Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004) is a 30-item version of the Youth Outcome Questionnaire-64 (YOQ-64; Burlingame et al., 2005). It is a brief psychotherapy outcome measure and maintains the parent measure's good psychometric properties (e.g., reliability, validity, and sensitivity to change). Its items are particularly sensitive to change (Berrett, 1999) and they tap into general symptoms relevant to many disorders and issues specific to youth. These characteristics make the YOQ an appropriate measure for tracking youth outcome over time.

The YOQ is a parent- or guardian-completed questionnaire for youth ages 4–17, with the option of being a self-report measure for youth who are 12 years or older. Items inquire about the past week of functioning and are written in first person at a 4^{th} grade level (e.g., "I have headaches or feel dizzy," "I steal or lie," "I feel irritated"). Parents completing the measure are

instructed to substitute the first-person "I" statements with "My child…" There are reports that parents are more effective in reporting externalizing behavioral problems (Murphy & Jellinek, 1990) and that youth are more effective in reporting internalizing phenomena (Merrell, 2001; Pagano, Cassidy, Little, Murphy, & Jellinek, 2000). Nonetheless, the YOQ has demonstrated high internal consistency reliability, sensitivity to change, and sensitivity and specificity in distinguishing clinical from nonclinical samples regardless the respondent (Burlingame et al., 2004).

The YOQ requires 5 minutes for completion. Its 30 items use a 5-point Likert-type scale and summative scoring to produce a total score for overall distress. Total scores may range from 0 to 120, with higher scores indicating greater distress. Scores at or above the established clinical cutscore of 29 (or 30 for self-report; Jacobson & Truax, 1991) are considered in the clinical range for distress levels. The reliable change index (RCI; Jacobson & Truax, 1991) for the YOQ is 10, indicating that score changes of 10 points or more represent true change and are distinguishable from measurement error. The YOQ has demonstrated an internal consistency reliability of .96. It has also demonstrated a concurrent validity of .76 with the Child Behavior Checklist (CBCL; Achenbach, 1991). Estimates suggest the YOQ has a four-week test-retest reliability of .83 (Burlingame et al., 2005).

Analyses

Analyses for this study were in two parts. The first part developed change trajectories for YOQ scores over time, identifying any variables predictive of these expected change trajectories. These trajectories inform the research literature as to what patterns of change may be expected and what variables seem to have an impact on these patterns. Similar trajectory models played an integral role in the second part of this study which tested the accuracy of a warning system

designed to identify clients at risk for negative outcome. Similar to past studies described above, these predictions were based on how the scores compare to prediction intervals around expected trajectories.

Creation of YOQ Change Trajectories

This study will use individual growth modeling—a type of multilevel modeling (MLM)—to create expected change trajectories for YOQ scores over time (R software, version 2.9.1, lmer model of lme4 package, full maximum likelihood estimation; SAS 9.2, mixed procedure, full maximum likelihood estimation; Singer & Willett, 2003). MLM is a form of regression that can be used to predict a client's score at any particular time (dependent variable) using a number of independent variables, among which is included a *time variable* (e.g., weeks or sessions in treatment). MLM estimated the intercept and slope of clients' YOQ score trajectories, which parameters constituted the fixed effects of the model. The model allowed these intercepts and slopes to vary randomly, also calculating variances related to each, which constituted the model's random effects.

The mixed (i.e., fixed and random) effects of individual growth modeling are not its only advantage over other longitudinal analysis techniques such as repeated measures regression. For example, MLM is effective even if data are collected at different intervals per client or if some measurement occasions have missing values. The longitudinal data (3 or more data points per client) that MLM uses also facilitates examination of more than just linear trajectory shapes (e.g., curvilinear or disjoint, using appropriate variable transformations and model parameters; Singer & Willett, 2003, pp. 208–213). This would be impossible using only two data points, as is the limit with pre- and post-treatment data. For example, in many other change trajectory studies (Bybee et al., 2007, Cannon et al., 2010; Finch et al., 2001; Warren et al., 2010) the best fitting

trajectory was curvilinear according to fit indices such as the -2 Log Likeliood or Bayesian Information Criterion (-2LL, BIC; Singer & Willett, 2003, pp. 208–213). These studies typically achieved curvilinearity by means of a natural log transformation of the time variable.

Variability in YOQ scores. MLM enabled us to calculate YOQ score variabilities at various levels. For example, we estimated the within-persons variance because of the expected correlation between scores that were nested within persons (i.e., repeated measures). We also estimated between-persons variances in intercept and slope, which were at a higher level in the model. We used an additional model to estimate variances within- and between-therapists, expecting that clients nested within therapists could have correlated scores. Considering the possibility for clients nested within treatment sites to have correlated scores, we also estimated variances within- and between-sites.

Predictor variables. This study's hypothesized individual growth model predicted YOQ scores (i.e., the dependent variable) using a time variable as well as by a handful of other independent variables. We tested various time variable transformations to determine which transformation fit the data the best according to fit indices such as the –2LL and BIC. The transformations tested included those from Mosteller and Tukey's ladder of powers (1977; e.g., square root and log transformations) as well as polynomial transformations (e.g., sessions + sessions^2 + sessions^3). Our plan was to use the best fitting transformation of either a sessions variable or a weeks variable as the time variable for the remainder of the study's models.

This time variable was useful for predicting scores over time, but we tested additional predictor variables as well. We tested dummy variables (0 = "no" 1 = "yes") for recent treatment (day treatment or inpatient treatment within 90 days of the start of the current outpatient episode), nonrecent treatment (90+ days in the past), and female. We tested continuous variables

for age, total number of sessions, total number of weeks, total number of YOQs, and mean number of sessions per month. Our hypothesized model tested all these predictors simultaneously, as both main effects (influencing trajectory elevation) and in interaction with the time variable (influencing trajectory slope or rate of change). We used a process of stepwise deletion of nonsignificant predictor variables from this hypothesized model to create a more parsimonious model. We then compared the predictor variables remaining significant in the model to the predictor variables of a model we created using a stepwise addition approach. After several subsequent iterations exploring the relationship of various variable combinations, we settled on an apparently optimal collection of variables for the final model.

Differences by initial severity. Some studies have addressed the correlation between trajectories' initial scores and rates of change (e.g., Cannon et al., 2010, Warren et al., 2010). These studies included initial score as a predictor in the model in efforts to control for the effects of all possible covariates to the independent variables of interest. For example, Cannon et al. (2010) examined trajectory differences by respondent (i.e., self- vs. parent-report) and included a covariate for initial score to ensure that differences perceived between the two respondent types were not actually attributable to systematic differences of initial severity between the two. Warren et al. (2010) also used this approach in their study examining trajectory differences in community mental health versus managed care settings. In additional approaches, these researchers tested samples from each setting that were matched by initial score and also tested for setting differences in a model that omitted any attention to initial score.

In contrast to the studies mentioned above, it would not have been appropriate for the present study to include initial score in its model predicting YOQ trajectories. This portion of the present study had the purpose of identifying predictors that were independent of the YOQ scores

themselves and the study did not examine any particular variable of interest. Whereas all other predictors have their origins external to the YOQ scores, initial score as a predictor has its origins from within the scores. Inclusion of the initial score predictor could have undesirably masked the extent to which other variables predict YOQ change trajectories, thus confounding the results; it would be a scenario in which one part of the dependent variable was used to predict another part of the same dependent variable. For these reasons, initial score was not examined as a predictor in the model.

Variable centering. To facilitate interpretation and reduce multicollinearity (Cohen, 2003, section 7.2; Singer & Willett, 2003, pp. 113–116), we centered continuous predictor variables around their grand means (e.g., $age - \overline{age}$). Mutlicollinearity refers to instances of high correlation between predictor variables that can result in instable estimates and inflated standard errors in regression models. Its confounding effects to interaction terms in a model can be overcome in part by centering predictor variables. To explain how a variable is centered, consider an example of subtracting the grand mean for age from the value of each client's age variable. This centering procedure would result in average aged clients having values near zero for their age variable (centered), older clients having positive values, and younger clients having negative values.

The more apparent benefit of variable centering is how it can facilitate interpretation of a model's estimates. Note that model estimates for intercept and slope correspond to a clients having zero as the value for all other predictor variables. However, zero is a very uncommon value for most predictor variables used in this study's models. For example, it would have been inconvenient for estimates of intercept and slope to correspond to clients aged zero or having zero total sessions. For centered variables, on the other hand, a zero value corresponds to the

mean for that variable (e.g., mean age or mean number of total sessions). The estimates in a model using centered predictor variables correspond to clients with average values for these predictors. This typically yields more intuitive interpretation of model estimates.

Model creation. This section reviews the creation of individual growth models in more detail. MLM produced multi-level models in which the Level 1 model predicted YOQ scores for any given individual. Using the notation conventions of Singer and Willett (2003), the basic equation representing this Level 1 model was

$$Y_{ij} = \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij},$$

where Y_{ij} is the YOQ score for person *i* at time *j*, π_{0i} is the trajectory intercept for person *i*, π_{1i} is the trajectory slope for person *i*, $TIME_{ij}$ is the value of a predictor variable indicating time in treatment (i.e., number of sessions or weeks) for person *i* at time *j*, and ε_{ij} is the within-person residual (i.e., error variance) or amount the observed score for person *i* at time *j* differs from predicted. In this Level 1 model, the π parameters are the fixed effects and the ε_{ij} parameter is the random effect.

The individualized intercept and slope parameters for each person's Level 1 model were predicted by Level 2 submodels that incorporated various independent variables. For example, a Level 2 submodel predicting the intercept parameter π_{0i} using age as a predictor variable would have the equation

$$\pi_{0i} = \gamma_{00} + \gamma_{01} AGE_i + \zeta_{0i},$$

where γ_{00} is the mean intercept (for clients with an average age, because *AGE* is centered), γ_{01} is the amount that the intercept differs per every unit that the individual's age exceeds the mean, and ζ_{0i} is the amount by which the observed intercept for person *i* differs from predicted. The

corresponding Level 2 submodel for the slope parameter π_{1i} would be very similar, having the equation

$$\pi_{1i} = \gamma_{10} + \gamma_{11} AGE_i + \zeta_{1i} .$$

This model's interpretation closely parallels the interpretation of the model for the intercept parameter, except that its parameters deal with slope rather than intercept. Examples aside, Level 2 submodels included the multiple predictor variables mentioned above, testing their effects on intercept and slope.

In these Level 2 submodels, the γ parameters represent the fixed effects and the ζ parameters represent the random effects. If a fixed effect estimate for a predictor variable such as AGE_i was statistically significant in the model, the implication was that age is systematically related to differences in change trajectory. Comparing the residual variances of a model that includes AGE_i to the residual variances of a model that does not include AGE_i indicates the percentage of variability accounted for by age (e.g., comparing the between-persons Level 2 variabilities in intercept or slope from each model, or the Level 1 within-person residual variabilities from each model).

Table 13 lists the example Level 1 and Level 2 models, along with the composite model they form once combined. Table 13 is merely an example using AGE_i as a predictor. The models that this study tested also included the other predictor variables mentioned above. Each parameter from the Level 1 Model can be substituted with the Level 2 submodel by which its value is predicted, creating an overall composite model. The last equation listed in Table 13 is an algebraic reformulation of the composite model. Its first two parameters ($\gamma_{00} + \gamma_{01}$) produce the trajectory intercept. Its next two parameters ($\gamma_{10} + \gamma_{11}$) produce the trajectory slope. The final

three parameters ($\zeta_{0i} + \zeta_{1i} + \varepsilon_{ij}$) enclosed in parentheses produce the random effects for the intercept, slope, and within-person residual, respectively.

Table 13

Examples of Level 1, Level 2, and Composite Models

Level	Model
Level 1	$Y_{ij} = \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij}$ (predicting trajectory using intercept and slope parameters)
Level 2	$\pi_{0i} = \gamma_{00} + \gamma_{01} AGE_i + \zeta_{0i}$ (predicting the intercept parameter from Level 1)
	$\pi_{1i} = \gamma_{10} + \gamma_{11} AGE_i + \zeta_{1i}$ (predicting the slope parameter from Level 1)
Composite	$Y_{ij} = (\gamma_{00} + \gamma_{01} AGE_i + \zeta_{0i}) + (\gamma_{10} + \gamma_{11} AGE_i + \zeta_{1i}) \times TIME_{ij} + \varepsilon_{ij}$
	$= \gamma_{00} + \gamma_{01} AGE_i + \gamma_{10} TIME_{ij} + \gamma_{11} AGE_i \times TIME_{ij} + (\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij})$

The hypothesized model nested scores within clients and within therapists. The nesting within therapists added two Level 3 submodels predicting intercepts and slopes for individual therapists. Each of these two Level 3 submodels would include a parameter for the mean intercept or slope, and an error term (i.e., random effect) for how the particular therapist's mean or slope differs from the overall average intercept or slope. The addition of these two random effects was the only modification required for the composite model when scores were also nested within therapists.

Warning System Prediction Accuracy

The second part of this study tested the accuracy of a warning system in its predictions of which clients would experience negative outcome. We used a split-samples approach to

create, and subsequently test, the accuracy of cutoffs intended to identify which ongoing outcome scores reached a severity predictive of negative final outcomes. We created these cutoffs using two primary approaches, and then tested how manipulating several features of these cutoffs impacted prediction accuracy.

Reference and validation samples. Our warning system based its predictions on outcomes observed in a reference sample comprised of half the 1,744 clients in our Part 2 data sample. (To note again, our Part 2 sample was a subset of the 4,309 clients in the Part 1 sample, selecting only clients with 3 or more YOQ measurements.) We tested the accuracy of these predictions in a validation sample comprised of the other half of the Part 2 sample. We created these two subsamples by random assignment. Usage of two separate subsamples attempted to avoid inflated estimates that could result from predictions being created from and tested on a single sample. To exercise additional caution, we performed the analyses of prediction accuracy ten times, each iteration using different random samplings, and reporting the mean of these various results.

Outcome class. The warning system attempted to predict which clients would experience negative outcome. A negative outcome corresponds to the deterioration outcome class. We determined the deterioration class and other outcome classes using the same two-step process used in similar past studies. Each of the two steps used cutoffs to evaluate different characteristics YOQ scores. The first step compared clients' overall YOQ change scores (i.e., difference between first and last YOQ scores) with the YOQ's reliable change index of 10 (RCI; Jacobson & Truax, 1991). The RCI is an index of the minimum amount of score change that is still distinguishable from measurement error. Clients whose change scores met or exceeded the cutoff of 10 points were those that we considered to have reliably changed.

Our second step in creating outcome classes compared the final YOQ raw score to the YOQ's clinical cutoff score of 29 (or 30 for self-report), identifying whether that final score fell within the clinical range. Thus we used our change score cutoff and our clinical cutoff to determine outcome classes. These outcome classes were *deterioration* if the final score was at least 10 points worse than baseline and in the clinical range (i.e., above the clinical cutoff), *no reliable change* if the final score differed from baseline by less than 10 points, *improvement* if the final score was at least 10 points better than baseline and above the clinical cutoff, or *recovery* if the final score was 10 points better than baseline and below the clinical cutoff. Clients whose final scores were at least 10 points worse than baseline but remained below the clinical cutoff at treatment termination fell in a subclinical form of the deterioration outcome class. The warning system described in the next section used nearly identical change score and clinical cutoffs to predict which clients were at risk for negative outcome.

Warning system cutoffs. This study's warning system monitored clients' ongoing YOQ scores during treatment, attempting to identify clients at risk for negative outcome by comparing clients' YOQ scores to the change score and clinical cutoffs described above. As to the latter cutoff—the clinical cutoff—we never allowed the system to signal a client as at risk for deterioration if the raw score for the most recent YOQ on record was below the clinical cutoff. Such scores were not even in the clinical range, were qualitatively different, and were thus of less concern. Although this clinical cutoff was in place for the whole of the study, we refer to it very little through the remainder of the study because our research focus was on the creation of the former cutoff, the change score cutoff.

We explored two main approaches to creating the former cutoff, whose purpose was to signal whether clients' ongoing scores were worsening by an amount large enough to be of

concern. One approach applied the cutoff to clients' change scores over time, whereas the other approach applied the cutoff to clients' raw scores over time. The remainder of this research report on these two approaches will refer to cutoffs based on change scores versus raw scores, yet both of these refer to ways of evaluating the magnitude of YOQ change scores; neither should be confused with the clinical cutoff score. Before describing the details of how we created these two types of cutoffs, we first provide a conceptual description of how our cutoffs functioned to identify clients at risk for negative outcome.

The warning system makes its predictions of negative outcome under the rationale that score deviations during treatment are predictive of final outcome. For example, a client whose midtreatment change score falls at the 95th percentile is showing rather severe negative ongoing outcome because higher YOQ raw scores—and change scores—indicate greater distress. This client is likely to have a final change score at or near the 95th percentile. Furthermore, if 10% of clients were expected to have final change scores showing reliable worsening (i.e., final scores 10+ points worse than baseline), then clients with final change scores above the 90th percentile (i.e., in the most extreme 10%) would presumably have reliably worsened. It follows then that midtreatment change scores at or above the 90th percentile would likely be predictive of clients at risk for reliable worsening. Such change scores associated with raw scores above the clinical cutoff could be predictive of clients at risk for deterioration.

Following this rationale, the warning system makes its predictions by comparing change scores at any given point in treatment to percentile rankings corresponding to that particular moment in treatment (e.g., percentile rankings for that particular session number). For an expected 10% of clients expected to have change scores that reliably worsened, the warning system would signal clients as at-risk if their change scores at any particular moment in

treatment were at or above the 90th percentile. We used percentiles in this way in this study, but we did not calculate these percentiles directly. Rather, we inferred these percentiles from a *t*-type confidence interval created around a modeled trajectory of expected change scores. We will also refer to such intervals as prediction intervals. This was a model of predicted YOQ change scores, whereas the change trajectories in Part 1 of this study were based on raw YOQ scores. The change scores were a measure of how much a client's scores differ from a personal baseline score and were calculated by recentering clients' raw scores around their respective baseline scores. As a result, the first score for each client was zero and subsequent scores indicated change from baseline. For example, a client with a baseline of 80 and subsequent scores of 75 and 72 would have had change scores of 0 (the baseline), -5, and -8.

The prediction intervals identified a set of change scores over time that served as the typical boundary between clients that had final outcomes in the deterioration outcome class and clients that did not. Change scores at any session that surpassed the boundary indicated that the client was at risk for negative outcome (e.g., deterioration, if the recent raw score was in the clinical range). Ultimately, these change score boundaries or cutoffs for deterioration and improvement could be displayed in a single reference chart, enabling clinicians to identify predicted final outcome given their client's session number and current change score. Figure 1 demonstrates an example of how such a chart could be constructed. To provide an example of how this chart uses ongoing change scores to predict final outcomes, the warning system predicts that clients with fifth session change scores of 13 (i.e., 13 points worse than baseline) will have final outcomes of deterioration. As another example, the warning system predicts that clients with fifth session change scores of 5 will have final outcomes of no reliable change. As a final

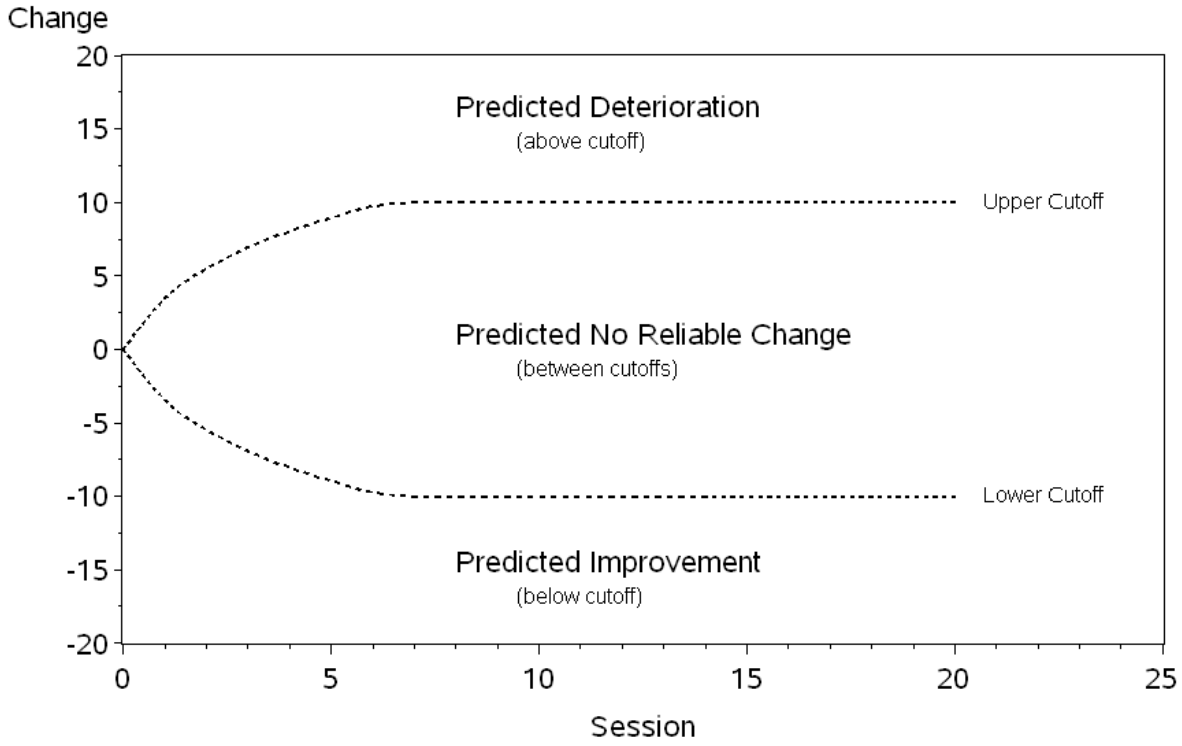

Figure 1. Example reference chart for predicting final outcome based on change score at any given treatment session. This chart is based on hypothetical data and is not intended for actual use.

example, the warning system predicts that clients with fifth session change scores of -13 will have final outcomes of improvement.

Prediction intervals in some past warning system studies (Bybee et al., 2007; Finch et al., 2001) have been 80% two-tailed intervals, which capture the center 80% of clients. The upper and lower boundaries of these intervals isolate the highest and lowest 10% of clients, the highest 10% corresponding to the 10% deterioration rate expected in these studies (Lambert & Bergin, 1994). Other studies have used prediction intervals based on deterioration rates observed in their specific sample. For example, Cannon et al. (2010) observed a deterioration rate of 16.4% and

thus calculated a 67.2% confidence interval in order to isolate the highest and lowest 16.4% of clients at any particular moment in treatment. In the present study we took a similar approach by calculating prediction intervals based on the percentage of clients in the reference sample who we observed to have reliably worsened change scores.

Whereas the target percentage of some past studies' prediction intervals was based exclusively on the reference sample's percentage of deteriorators (i.e., clients who reliably worsened and had a final score in clinical range), we based our prediction interval's target percentage on the percentage of clients in the reference sample whose change scores reliably worsened, regardless whether clients' final scores fell in the clinical or subclinical ranges. Our rationale in this methodological departure was to have the two steps of predicting deterioration more strictly observe the existing distinction between the previously established two steps of determining actual deterioration. The first step of determining actual deterioration examines change score magnitude for whether it qualifies clients as candidates for deterioration. Clients demonstrating sufficiently large worsening are only candidates; they are not considered actual deteriorators until the second step of the determination process confirms that their final YOQ score is in the clinical range.

Similarly, our first step of predicting deterioration used prediction interval cutoffs to identify candidates for deterioration; that is, all clients whose change scores showed sufficient worsening, and who might be predicted to deteriorate if in the next step they are shown to have most recent scores in the clinical range. Thus we considered it appropriate for these cutoffs to have a target percentage corresponding to all candidates: the combination of clinical and subclinical deteriorators. Had the target percentage that we created from the reference sample omitted the subclinical deteriorators, it could have underestimated the actual percentage of

clients in the validation sample with change scores making them candidates for deterioration. This is because our simulation of applying the warning system in a clinical setting did not permit us to remove from the validation sample the clients who would go on to become subclinical deteriorators; such clients would not be identifiable midtreatment, when the system would be applied. Thus the target percentage would be created from only a subset of the type of clients it was trying to identify.

Once the warning system's step one cutoffs identified candidates for deterioration based on change scores, the second step of predicting deteriorators then determined which candidates to signal as at risk for deterioration based on whether the most recent YOQ raw scores fell in the clinical range. Although our study did not focus on this second step of evaluating raw scores, such evaluation is critical for interpreting symptom severity, predictions of deterioration, and final classifications of deterioration. A warning signal would likely be of less concern, or even common, for a client whose baseline raw score was in the subclinical range. In contrast, a signal would likely be more alarming for a client whose baseline was very high in the clinical range, and who would thus be expected to have significantly reduced scores over time.

We created only one prediction interval or set of cutoffs for change scores because the criteria for deterioration were universal (i.e., an increase of 10 points or more for any and all clients). If the YOQ were to have criteria for deterioration that differed by subpopulation, it would be appropriate to have prediction intervals or cutoffs specific to each subpopulation. However, the deterioration criteria are universal regardless clients' individual differences (including initial score) and thus we calculated only a single prediction interval and its corresponding single set of cutoffs for deterioration. In terms of MLM, this meant that we included no predictors other than the time variable in the change score model that is at the heart

of the warning system's prediction intervals and cutoffs. The time variable was necessary to create a nonzero slope for the model.

The approach described above of using change scores as the basis for prediction intervals and cutoffs differs from past studies' approaches of using raw scores as the basis (Bybee et al., 2007; Cannon et al., 2010; Finch et al., 2001). For example, instead of creating prediction intervals around change scores that always begin with zero (i.e., the recentered baseline), these past studies created prediction intervals around raw scores that could start with whatever the raw (i.e., uncentered) baseline score happened to be. The upper boundary of the prediction interval served as the cutoff for at-risk status and was represented by raw scores rather than change scores.

The cutoffs of these past studies had to accommodate clients' varying initial scores because whereas a client with a baseline score of 80 might have a fifth session cutoff of 89, a client with a baseline score of 50 would need a much lower cutoff. These studies would ideally have made models and prediction intervals for every possible baseline score, but they typically had too few data to create so many separate models. Instead, they stratified the data according to baseline score, splitting clients into brackets or *score bands*, and created separate models and prediction intervals for each. This score band approach was fairly successful in these past studies.

Figure 2 demonstrates an example warning system reference chart for cutoffs created using raw scores and score bands. The chart shows the expected raw score trajectory and associated cutoffs for the score band comprised of clients with baseline scores in the range of 47 to 53. To provide an example of how this chart uses ongoing raw scores to predict final

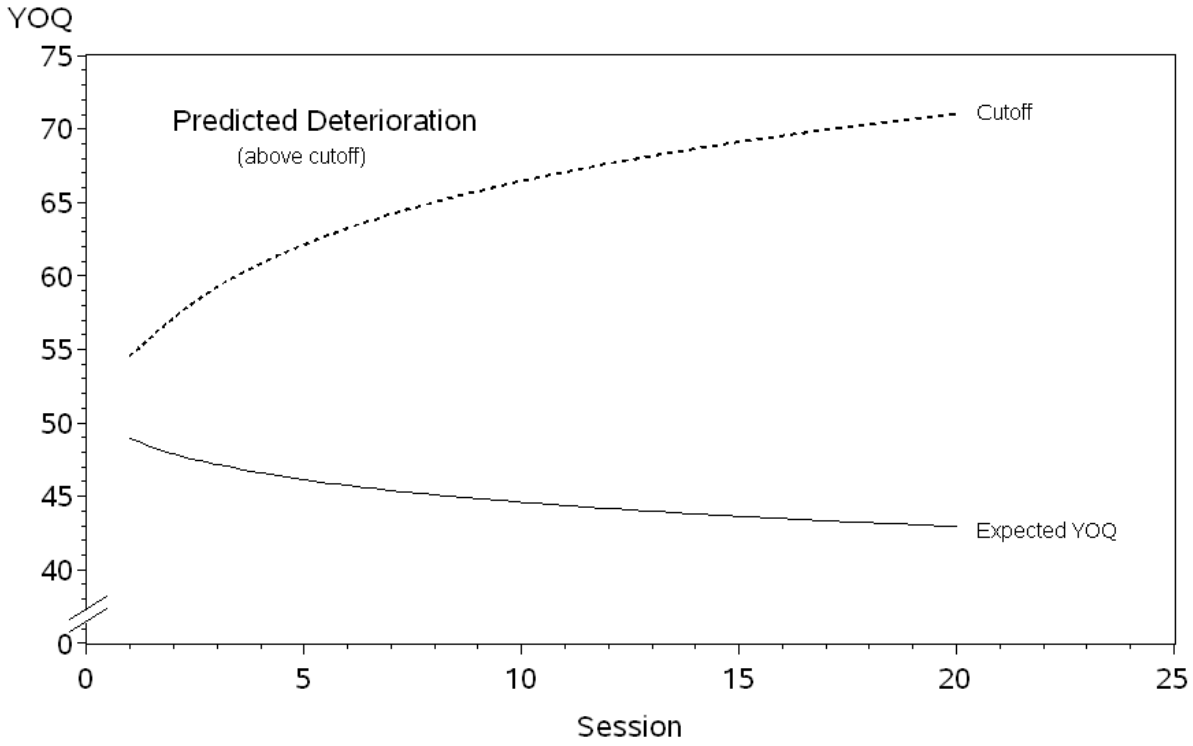

Figure 2. Example reference chart for predicting final outcome based on raw score at any given treatment session. This chart is for the score band comprised of clients with baseline scores in range of 47 to 53. This chart is based on hypothetical data and is not intended for actual use.

outcomes, the warning system predicts that clients with fifth session raw scores of 65 will have final outcomes of deterioration. As another example, the warning system predicts that clients with fifth session raw scores of 55 will have final outcomes of no reliable change. As a final example, the warning system predicts that clients with fifth session raw scores of 40 will have final outcomes of improvement.

In the present study we also tested warning system prediction accuracy using the score band approach to creating cutoffs and identifying clients at risk for negative outcome. We compared the prediction accuracy for the change score approach to the accuracy of the score

band or raw score approach. One potential advantage to the change score approach was the greater number of clients that were used in the model producing the prediction intervals. We could use all clients at once in a single model using the change score approach. In contrast, with the raw score approach we had to use only a portion of the overall clients per model because it had to create separate models per score band. This difference in sample size may have contributed to some differences in prediction accuracy we found the raw score and change score approaches. A second potential advantage to the change score approach could be the need for only a single reference chart for the warning system's outcome predictions, as opposed to separate charts for each score band of the raw score approach. Our primary evaluation of the change score approach, however, was based on its comparative accuracy in predicting which clients ultimately experienced negative outcome.

Compared to YOQ raw scores, YOQ change scores carry less information in that they do not account for symptom severity on an absolute scale, but only on a scale relative to each client's baseline. Allen and Yen (1979) demonstrated that difference scores (i.e., change scores) tend to be less reliable than the raw scores from which they are calculated. However, we anticipated that the problems of weaker reliability for change scores would have minimal impact on their use in this study. This study used change scores to predict other change scores, that is, ongoing midtreatment change scores to predict final change scores. It was those final change scores in comparison with the YOQ's RCI value of 10—yet another change score—that were the basis for the various outcome classes. As discussed and demonstrated throughout the Results section below, the baseline-related information lacked by the change scores we used to create our warning system cutoffs would likely not have added any benefit to the warning system prediction accuracy had it been present.

Warning system prediction accuracy. With cutoff scores established for which change scores and raw scores would signal clients as at risk for negative outcome, this study next calculated the warning system's prediction accuracy by comparing its outcome predictions to the actual outcomes observed in the data. We established the prediction intervals and cutoffs using the reference sample (i.e., subsample 1), then used these cutoffs to predict the outcomes of clients in the validation sample (i.e., subsample 2). Scores exceeding the cutoff on any occasion except the final measurement signaled clients as predicted to have final outcomes of deterioration. The study reported the accuracy of these predictions in a contingency table comparing predicted final status (i.e., deterioration vs. non-deterioration) to actual (i.e., observed) final status. This table identified the number of true positives, false positives, true negatives, and false negatives. The table facilitated calculation of the warning system's accuracy in identifying deteriorators. These calculations of accuracy included sensitivity (percentage of actual deteriorators correctly predicted), specificity (percentage of actual non-deteriorators correctly predicted), hit rate (percentage of predictions that were correct—of any type), positive predictive power (percentage of predicted deteriorators that are actual deteriorators), and negative predictive power (percentage of predicted non-deteriorators that are actual non-deteriorators).

We calculated separate prediction accuracies for the change score versus the raw score (i.e., score band) approaches for creating cutoffs as described above. We contrasted the accuracy and method of the cutoffs from these two approaches. In post hoc analyses, we manipulated various cutoff characteristics and calculated the corresponding prediction accuracies. We save our explanation of these characteristics for the Results section below, given the post hoc nature of their examination. We sense that these characteristics are better explained in the context of the

prediction accuracy results for our originally planned change score and raw score cutoffs. After examining prediction accuracies for various cutoffs, we suspected that our predictions were failing for clients whose final scores deviated from the general trend of their previous scores. We examined this possibility by plotting trajectory shapes in terms of a plotted point for baseline, a plotted point for the mean midtreatment change score, and a plotted point for the final change score. We created separate plots for clients that we correctly predicted as deteriorators (true positives), correctly predicted as non-deteriorators (true negatives), incorrectly predicted as deteriorators (false positives), and incorrectly predicted as non-deteriorators (false negatives).

RESULTS

The analyses for this study were in two parts. The first part developed change trajectories for YOQ scores over time, identifying variables that were predictive of the intercept and slope of these trajectories. Similar trajectory models played a role in the second part of this study. This second part tested the accuracy of a warning system designed to identify clients at risk for negative outcome based on how YOQ scores over time compared to prediction intervals around expected trajectories.

YOQ Change Trajectories

We used MLM to model YOQ change trajectories. The model's random effects enabled us to calculate YOQ score variabilities associated with differences between clients, therapists, and treatment sites. The model's fixed effects enabled us to quantify the relationship between predictor variables and change trajectories' intercept and slope.

Variability in YOQ Scores

Multilevel modeling produces estimates for fixed effects and random effects. The random effects are a measure of variability the model's predictors have not explained. We used a model with no explanatory variables to demonstrate how variability in YOQ scores was distributed among clients and therapists. Such a model is called an *unconditional means* model; its only fixed effect parameter is a constant for the YOQ trajectory intercept, the estimate of which is simply the overall mean YOQ score (40.2), with no conditions (i.e., predictors). The unconditional means model with YOQ scores nested within clients and within therapists produced random effects estimates for clients and therapists. We tested the statistical significance of these parameters one at a time by identifying the deviance statistic for the model with and the model without the parameter in question. We then calculated the difference in these

two deviance statistics and compared this value with the .05-level critical value on a chi-square distribution (Singer and Willett, 2003, explain significance testing using the deviance statistic).

For example, compared to a model with scores nested only within clients, a model with scores also nested within therapists had a deviance 85 units lower (106412 - 106327 = 85). This value exceeds the .05-level critical value of 3.84 on the chi-square distribution for 1 degree of freedom; there was only one parameter different between these models. We also tested a model with an additional parameter for nesting scores within treatment sites, but the deviance statistic remained unchanged, indicating that these data appeared to have no variability attributable to site, while controlling for variability attributable to client and therapist. This finding may likely be a result of the limited variance in site given that 97.4% of YOQs were administered at just one of the 9 sites on record.

Given these results, the most appropriate nesting of YOQ scores appeared to be within clients and within therapists. The majority of the variability in scores was associated with variability between clients—64% (variance = 200.18)—whereas 29% (variance = 89.77) was associated with variability within clients (each client's scores on one occasion to the next) and 7% (variance = 21.68) was associated with variability between therapists. As reported in Table 3, 89% of clients had only a single therapist (i.e., were fully nested within therapist) and 98% had only a single site on record. Whereas these numbers account for therapists and sites associated with treatment sessions at which no YOQ was recorded, the MLM random effects only accounted for the therapists and sites associated with each YOQ measurement, not fully capturing the effects of variation in therapist or site between measurement occasions. For example, rates of being fully nested within therapist and site were higher when examining only YOQ measurement occasions (93% within therapist and 99% within site).

We next examined an *unconditional growth model* to determine the portion of variability in YOQ score trajectories' elevations attributable to clients versus therapists, as well as a similar breakdown in variability in trajectories' slopes. The unconditional growth model included a single parameter accounting for time across which YOQ scores were observed, with no other conditions (i.e., predictor variables) affecting trajectory growth (i.e., slope). As will be explained below, the time variable we selected was the natural logarithm of session number (*LNSESS*). We included the *LNSESS* variable as both a fixed effect and a random effect in the model, the latter effect modeling YOQ trajectory slopes as varying at random and producing an estimate of the associated variance for both clients and therapists. Eighty-four percent of the variability in trajectory slopes was associated with differences between clients (variance = 24.27), versus 16% that was associated with differences between therapists (variance = 4.76). Ninety-two percent of the variability in trajectory elevations was associated with differences between clients (variance = 265.48), versus 8% associated with differences between therapists (variance = 21.62). Trajectory intercepts (a measure of trajectory elevation) were correlated with trajectory slopes at $r = -.61$.

Predictor Variables

We examined a number of predictor variables for their relationship to change trajectories' intercepts and slopes. We began this examination process by creating a model that included all these predictors simultaneously, as both main effects (influencing trajectory elevation) and in interaction with the time variable (influencing trajectory slope or rate of change). We describe below the various steps we took in reducing the hypothesized model down to on an apparently optimal collection of variables for the final model.

Hypothesized model. Table 14 presents estimates for the variables we hypothesized would likely be significant as predictors, or fixed effects, in the multilevel model. We used the hypothesized model in Table 14 as somewhat of a starting point and basis for creating our final model. The estimates in the first column of the table are related to trajectory intercepts, or elevations. The first estimate listed is for Intercept and indicates that the modeled baseline YOQ score was 43.8 (Table 14, row = Constant, column = Intercept). The model produced this estimate while controlling for the effects of the other variables in the model. This estimate corresponds to clients with values equal to zero for the other predictor variables in the model. A value of zero corresponded to "no" for dummy variables such as prior treatment (0 = "no", 1 = "yes") or to the mean value for continuous variables such as age. Continuous variables were centered around their mean, as explained in the Method section.

The estimate appearing next in the first column in Table 14 indicates that trajectories for clients with prior treatment (nonrecent: at least 90+ days in the past) were typically 2.4 points higher, while controlling for the effects of all the other predictor variables (i.e., their values being equal to zero). Similar interpretation applies to the remaining estimates in the first column of the table. One variation was that the main effect for total number of sessions was not statistically significant without interaction with the *LNSESS* time variable, as will be discussed below. In brief, intercepts were much higher for clients with prior treatment within the past 90 days (i.e., transitioned to the outpatient setting from the day treatment or inpatient setting), higher for older clients, higher for clients who ended up having more sessions per month, yet lower for clients having more YOQ measurements per month, and slightly lower for female clients.

The estimates in the second column of Table 14 are related to trajectory slopes, or the rate of change in YOQ scores over time. These estimates for slope are expressed in units

Table 14

Hypothesized Change Trajectory Model

	Intercept		Slope (interaction w/ *LNSESS*)	
Fixed Effects	Estimate	*SE*	Estimate	*SE*
Constant	43.77*	0.45	-3.27*	0.25
Prior treatment (90+ days in past)	2.42*	0.79	1.00*	0.43
Prior treatment (within past 90 days)	11.04*	2.33	-2.05	1.20
Age	0.72*	0.11	-0.18*	0.06
Total no. of sessions	0.12	0.07	0.09*	0.03
No. sessions per month	0.76*	0.37	-0.16	0.20
No. YOQs per month	-1.52*	0.64	-0.51	0.42
Female	-1.90*	0.62	0.11	0.33

	Intercept		Slope (*LNSESS*)		
Random Effects	Estimate	*SD*	Estimate	*SD*	*r*
Between Clients	254.34*	15.95	23.68*	4.87	-.50
Between Therapists	20.29*	4.51	4.44*	2.11	-.40
Within Clients (residual)	69.77*	8.35	—	—	—

Note. $N = 4{,}309$. Estimates for the Constant parameter reflect the mean intercept and slope where other variables were equal to zero, corresponding to "no" for dummy variables (i.e., value = 0 for the two prior treatment variables and female) and corresponding to the grand mean of continuous variables (i.e., value = 0 for age, total no. of sessions, total no. of weeks, sessions per month, and YOQs per month; these variables were centered around their respective grand means). The other estimates are deviations from these constants. Dashes mark table cells where no estimate would be relevant.

*$p < .05$.

corresponding to our chosen time variable, *LNSESS*. As reported in the Method section above, we selected this time variable by testing a various transformations of the number of sessions and weeks that had passed in treatment at the time of each YOQ measurement. A natural logarithmic transformation of the sessions variable demonstrated superior model fit according to the deviance statistic. The transformation we selected was *LNSESS* = $\log_e$(sessions + 1). Where the sessions variable is equal to zero, *LNSESS* is also equal to zero; whereas the two variables begin equal, they differ over time. The transformed *LNSESS* achieves a curvilinear trajectory by decrementing the effect of sessions over time. Slopes begin steeply downward, corresponding to quick reduction in distress according to YOQ scores, but the slopes taper off over time.

As illustrated in Figure 3, the first estimate listed in Table 14 on the row labeled Constant and in the column labeled Slope rounds to -3.3 and corresponds to the change in YOQ scores per every one unit change in the *LNSESS* time variable, while controlling for the effects of the other variables. When *LNSESS* = 1, sessions = 1.7, so the model predicts YOQ scores to decrease by 3.3 points in the first 1.7 sessions. However, when *LNSESS* = 2, sessions = 6.4, which means that the subsequent drop of 3.3 points is predicted to require another 4.7 sessions (6.4 – 1.7 = 4.7). Continuing, where *LNSESS* = 3, sessions = 19.1; the next 3.3 point decrease requires another 12.7 sessions (19.1 – 6.4 = 12.7).

The next estimate appearing in the second column of Table 14 indicates that slopes were not as steep for clients with prior treatment (nonrecent: 90+ days in past), the rate of change being reduced by 1.0 points per one unit change in *LNSESS*. The figure created for the final model will provide further illustration of how slopes differed by predictor variable. Other slope-related parameters that were statistically significant in the hypothesized model showed that older clients had faster rates of change but clients with more sessions had slower rates of change.

Final model. The hypothesized model had several nonsignificant parameters, which suggested that a more optimal model could be found. Employing a process of stepwise deletion, stepwise addition, and various iterative models exploring relationships between variables, we produced a final model with all significant parameters, as presented in Table 15. Note that although the main effect for the variable indicating total number of sessions was not significant on its own, the interaction of this variable with *LNSESS* (i.e., its effect on slope) was significant. We retained the main effect in the model in order for the model to be hierarchically well specified (Peixoto, 1987, 1990).

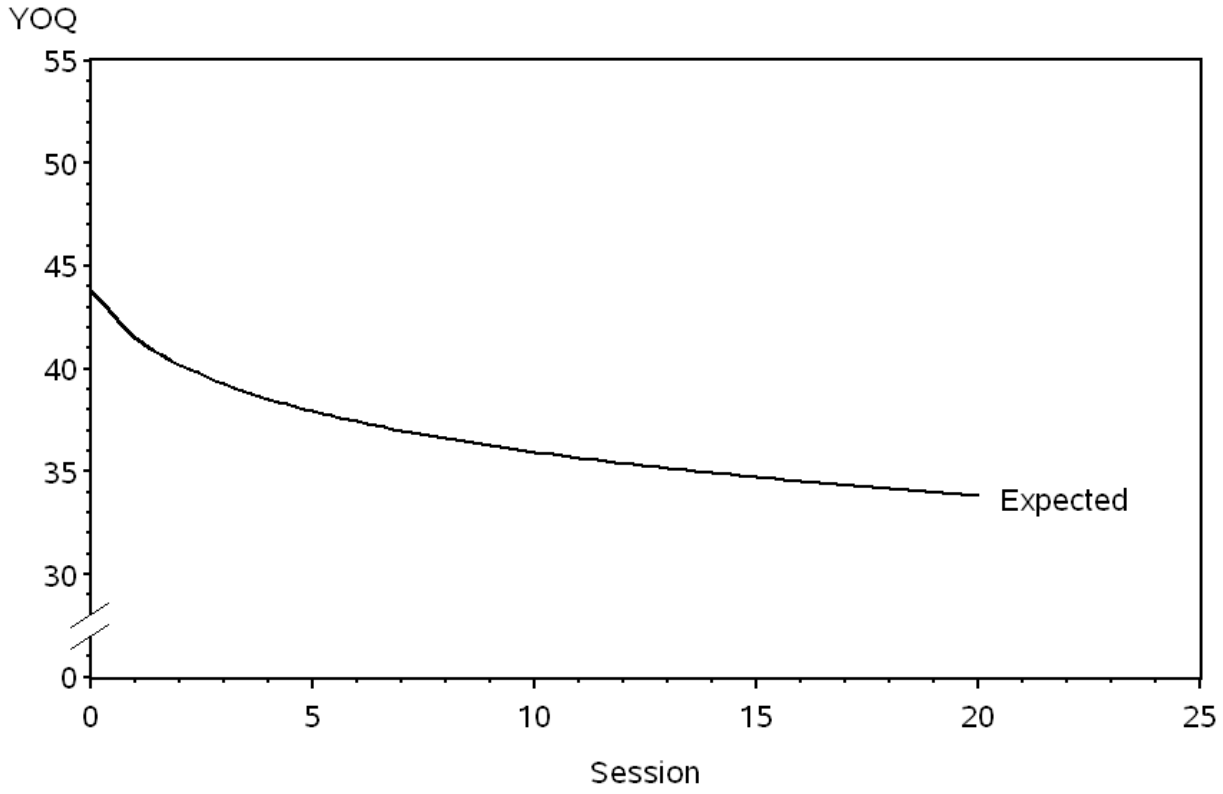

Figure 3. Curvilinear LNSESS time variable.

Table 15

Final Change Trajectory Model

	Intercept		Slope (interaction w/ *LNSESS*)	
Fixed Effects	Estimate	*SE*	Estimate	*SE*
Constant	43.77*	0.42	-3.27*	0.23
Prior treatment (90+ days in past)	2.31*	0.79	1.06*	0.43
Prior treatment (within past 90 days)	8.12*	1.63	—	—
Age	0.73*	0.11	-0.19*	0.06
Total no. of sessions[a]	0.02	0.08	0.08*	0.03
Total no. of weeks	0.05*	0.02	—	—
No. sessions per month	0.94*	0.33	—	—
No. YOQs per month	-1.51*	0.56	-0.72*	0.29
Female	-1.70*	0.46	—	—

	Intercept		Slope (*LNSESS*)		
Random Effects	Estimate	*SD*	Estimate	*SD*	*r*
Between Clients	255.02*	15.97	23.79*	4.88	-.50
Between Therapists	19.80*	4.45	4.37*	2.09	-.39
Within Clients (residual)	69.77*	8.35	—	—	—

Note. N = 4,309. Estimates for the Constant parameter reflect the mean intercept and slope where other variables were equal to zero, corresponding to "no" for dummy variables (i.e., value = 0 for the two prior treatment variables and female) and corresponding to the grand mean of continuous variables (i.e., value = 0 for age, total no. of sessions, total no. of weeks, sessions per month, and YOQs per month; these variables were centered around their respective grand means). The other estimates are deviations from these constants. Dashes mark cells where no estimate was calculated, either because of nonsignificance in the model (e.g., fixed effects) or because of irrelevance (e.g., random effects).

[a]The main effect for total number of sessions was retained in the model despite nonsignificance in order for the model to be hierarchically well specified (Peixoto, 1987, 1990).

*$p < .05$.

Table 15 shows that estimates for the constants for intercept and slope were essentially the same in the final model compared to the hypothesized model. Controlling for the effects of all other variables, the modeled baseline YOQ score was 43.8 and the rate of change was -3.3 points per one unit change in *LNSESS*. The sample producing this model was fairly large, bringing into question whether the statistical significance of some parameters was more attributable to the large sample size than to a notable effect size. Formal analysis of effect size for multilevel modeling is very complex and we chose the practical approach of visually inspecting how the different variations on the expected trajectory compared in Figure 4. Each trajectory depicted corresponds to a single predictor variable having a nonzero value while the other predictors remain at zero. The dummy variables are each shown as having a value of one. For example, when prior treatment (nonrecent: 90+ days in past) = 1, or "yes", this corresponds to a trajectory with an intercept that is 2.3 points higher and with a slope that is 1.06 points slower than the average (compare trajectories labeled "Nonrecent treatment" and "Expected" in the figure). The continuous variables are each shown as having a value one standard deviation above the variable mean (see Table 3 for *SD*s). For example, an additional YOQ per month corresponds to a trajectory with an intercept that is 1.5 points lower and with a slope that is 0.72 units faster. The figure depicts a trajectory for a client with the standard deviation of .842 more YOQs per month than the mean of 1.0 and the trajectory is noticeably lower than the average expected trajectory. Although the trajectory differences according to age, total weeks, total sessions, and sessions per month were statistically significant, Figure 4 demonstrates that these differences may be of little clinical significance.

The above examination of change trajectories adds to the research literature on factors associated with psychotherapy outcomes. The models presented above generally predict positive

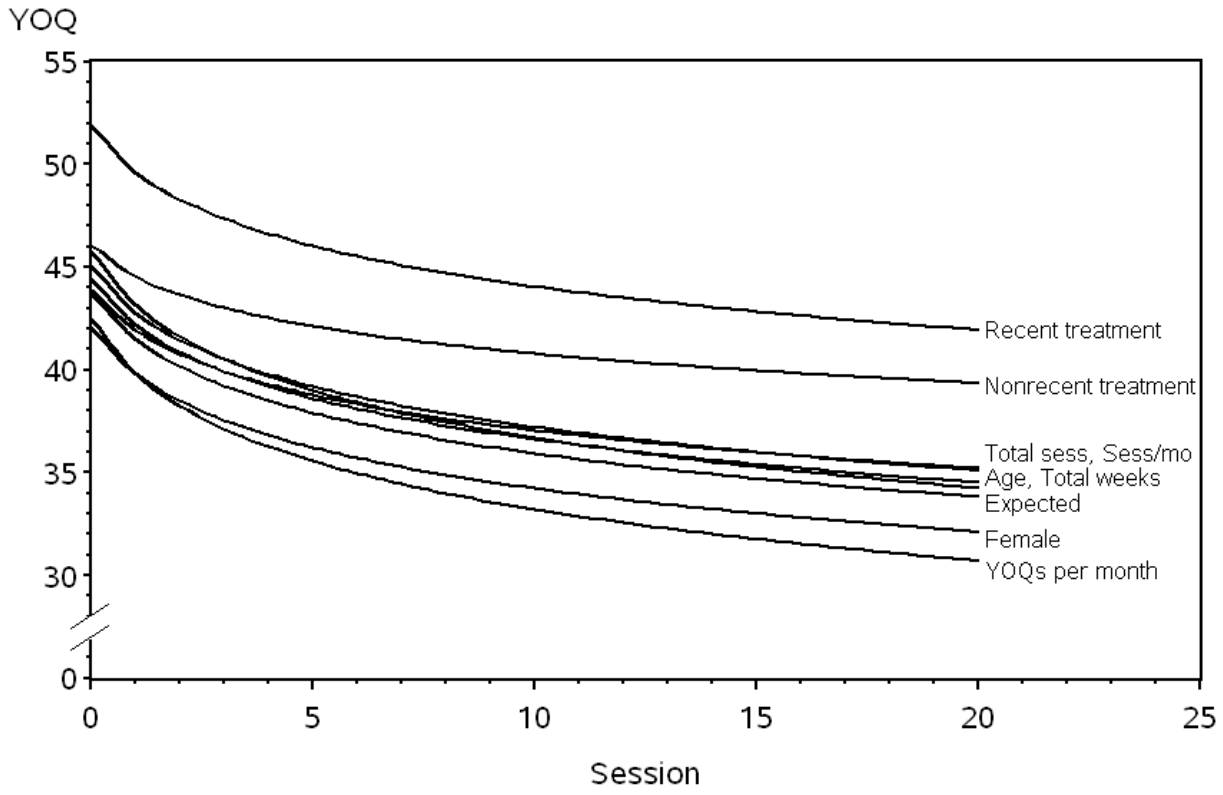

Figure 4. Various change trajectories accounted for in final model.

outcomes in terms of reduced distress scores as measured by the YOQ, roughly a 10-point reduction in 21 sessions of psychotherapy. As described in the Method section, a 10-point reduction is considered reliable change according to the YOQ's reliable change index (RCI, Jacobsen & Truax, 1991) of 10 points, indicating the minimum change in scores that is still distinguishable from measurement error.

Warning System Prediction Accuracy

For the second part of this study, we tested the accuracy of a warning system in its predictions of which clients would experience negative outcome. As we describe in more detail below, we created predictions of outcomes using a randomly selected half of the sample. We tested the accuracy of these predictions in the other half of the sample, calculating indices such

as the sensitivity and specificity of our predictions (alerts) for deterioration. The first subsample functioned as the reference sample and the second subsample functioned as the validation sample. We created the two subsamples using random assignment to avoid possible systematic differences between the samples that could confound the results. To further negate how this subsample creation may have influenced the results of this portion of the study, the results we present below are the mean results of 10 different random samplings.

Our predictions of clients who would have negative outcome in subsample 2 relied, in part, on our expectation of the percentage of clients to experience negative outcomes. We designed the warning system to identify a target percentage of clients corresponding to the percentage of clients with demonstrating reliable worsening in the reference sample. Table 16 presents the percentages of each outcome class in the reference sample.

Table 16

Outcome Classes for Part 2 Reference Sample

Outcome class	*n*	%
Recovery	128	14.6%
Reliable improvement	165	18.8%
No reliable change	456	52.1%
Deterioration	117	13.3%
Subclinical deterioration	10	1.2%

Note. $N = 876$.

Warning System Cutoffs

A primary purpose of the warning system was to identify clients whose YOQ scores were increasing, which typically corresponds to increased distress, and which put them at risk of

finishing treatment in the deterioration outcome class. We tested and compared two approaches to monitoring YOQ scores for such signaling increases. In our first approach, we examined clients' YOQ change scores over time (i.e., equal to raw score minus the client's baseline) and used a change score threshold based on a prediction interval as the cutoff for whether clients would be signaled as at risk for deterioration. In our second approach, we examined clients' raw YOQ scores over time and used a predetermined raw score threshold as a similarly functioning cutoff. Whether creating the cutoffs based on change scores or raw scores, we created the cutoffs in the same manner. We created multilevel models of the reference sample's change scores or raw scores over time. These models were unconditional growth models, the only predictor variable being a time variable *LNSESS*, as described and used in Part 1 above. Our modeling procedure also calculated a two-tailed *t*-type confidence interval around the predicted scores over time (using the ALPHAP = option of SAS PROC MIXED). We configured this prediction interval such that its upper boundary served as the cutoffs isolating the highest 14.5% of predicted scores. This percentage corresponded to the reference sample's percentage of clients whose scores reliably worsened over time (14.5% = 13.3% deterioration + 1.2% subclinical deterioration; see Table 16). We later used these cutoffs created from the reference sample to predict which clients in the validation sample would show deterioration. For cross-reference, Table 17 presents the outcome classes for the larger sample of used in Part 1 of this study.

By design, the change score baseline was equal to zero for all clients, necessitating only a single set of cutoffs over time. On the other hand, our cutoffs for raw scores had to account for the varying baselines. To do this, we stratified the reference sample data by baseline score, yielding 7 score bands. The sample size for score bands 1–6 ranged from 117 to 133, for both the reference and the validation samples. The sample size for score band 7 ranged from 99 to 104.

Table 17

Outcome Classes for Part 1 Sample

Outcome class	*n*	%
Recovery	546	12.7%
Reliable improvement	679	15.8%
No reliable change	2,486	57.7%
Deterioration	553	12.8%
Subclinical deterioration	45	1.0%

Note. $N = 4{,}309$.

The final score band had fewer clients because the process of creating the score bands attempted to select at least 120 clients per score band, starting its grouping process with clients having the lowest baseline scores and creating groups as it proceeded to clients with the highest baseline scores. The ten iterations of random sampling and inconsistent variability of baseline scores precluded perfectly even sample sizes for all score bands, with fewer clients being available for this final score band. Our experimentation with aiming to select slightly fewer than 120 clients per score band occasionally created an eighth score band, which would have introduced complications it was better to avoid. Returning focus to the purpose of cutoff creation, we created separate models for each score band, the corresponding prediction intervals or cutoffs thus accounting for variability in baseline scores.

Table 18 shows the baseline ranges for each score band in the reference sample. The table goes on to show the multilevel model estimates for intercept and slope for each score band. Note the expected difference in rate of change (i.e., slope) per score band. Higher baseline scores are associated with faster rates of change. The table also presents the specific YOQ scores expected after particular numbers of treatment sessions, along with the corresponding cutoff scores to

Table 18

Predicted Scores and Cutoffs for Score Bands and Change Scores

Score band	Baseline range	Model estimates			Baseline	Score after session no.								
		Intercept	Slope			1	2	3	4	6	8	10	15	20
				Cutoff		23	27	31	33	37	39	42	46	49
1	0–23	15.31	2.82	Expected	15	17	18	19	20	21	21	22	23	24
				Cutoff		33	37	40	43	46	49	51	55	57
2	24–31	27.43	.87	Expected	27	28	28	29	29	29	29	30	30	30
				Cutoff		39	43	46	48	52	54	56	60	63
3	32–38	36.23	-1.12	Expected	36	35	35	35	34	34	34	34	33	33
				Cutoff		48	51	53	55	58	60	61	64	67
4	39–46	43.71	-2.84	Expected	43	42	41	40	39	38	37	37	36	35
				Cutoff		55	57	59	61	63	65	66	69	71
5	47–53	50.67	-2.54	Expected	51	49	48	47	47	46	45	45	44	43
				Cutoff		64	66	67	69	70	71	72	74	76
6	54–63	59.94	-5.50	Expected	60	56	54	52	51	49	48	47	45	43
				Cutoff		82	83	83	84	85	85	86	87	88
7	64–120	75.78	-8.84	Expected	76	70	66	64	62	59	56	55	51	49
				Cutoff (unrestricted)		5	9	12	15	18	21	23	27	30
				Cutoff (restricted)		5	9	10	10	10	10	10	10	10
Change scores		0.79	-2.26	Expected	0	-1	-2	-2	-3	-4	-4	-5	-5	-6

Note. Model estimates (fixed effects) are all significant at the $p < .05$ level except the slopes for score bands 2 and 3.

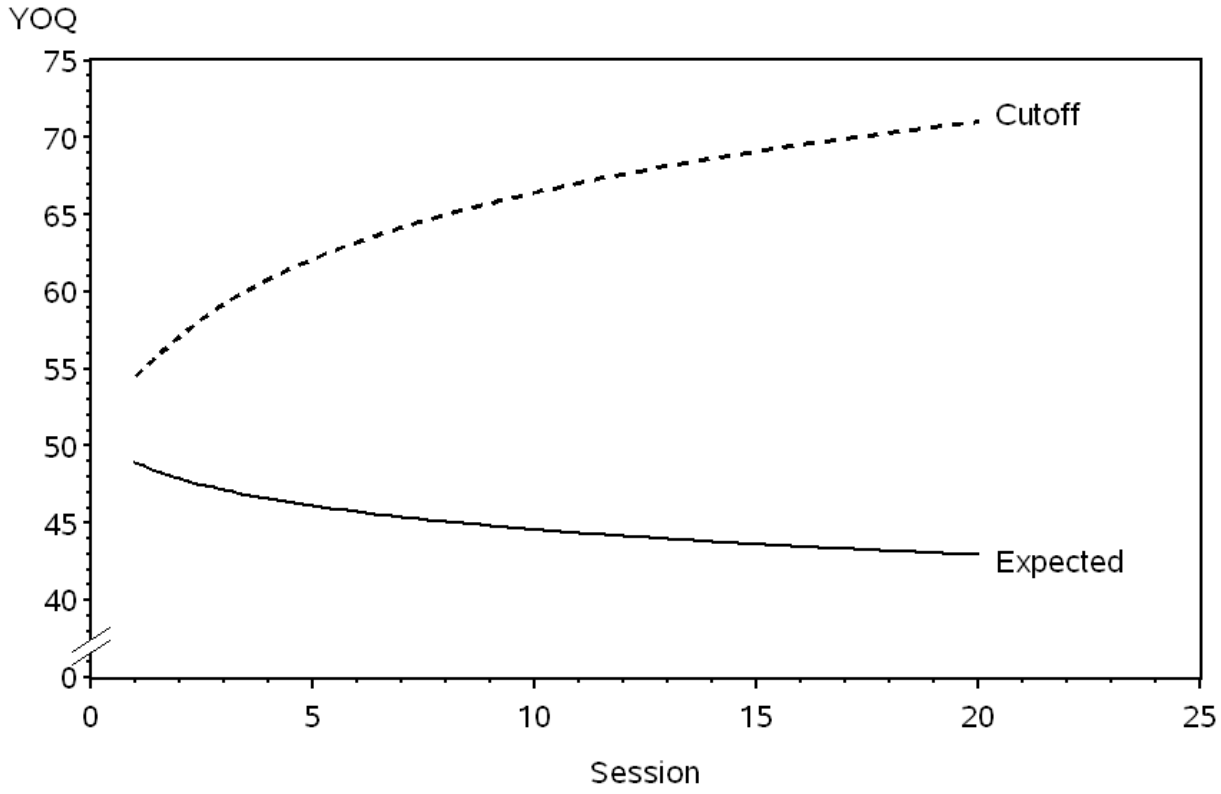

Figure 5. Predicted scores and cutoffs for score band 5.

signal clients as at risk for deterioration. Figure 5 illustrates an example of the expected scores and cutoffs corresponding to a baseline of 51. The cutoff expands upward over time given that it is merely the upper boundary of a confidence interval around the predicted scores. Were it shown in the figure, the lower boundary of the interval would mirror the upper boundary such that the two would expand out over time as prediction error increases toward the latter parts of the modeled trajectories.

Table 18 also presents intercept and slope estimates for a model of the reference sample's change scores over time. The table also shows the associated expected scores and cutoffs, as illustrated in Figure 6. The model predicted a mean decrease of 6 points after 20 sessions for the reference sample overall. If unrestricted, the associated cutoffs extend upward, similar to those

for each score band. We originally reasoned that such cutoffs could extend too high to be effective and thus planned to restrict the cutoffs to a maximum change score of 10 points. Figure 6 illustrates these restricted cutoffs as well.

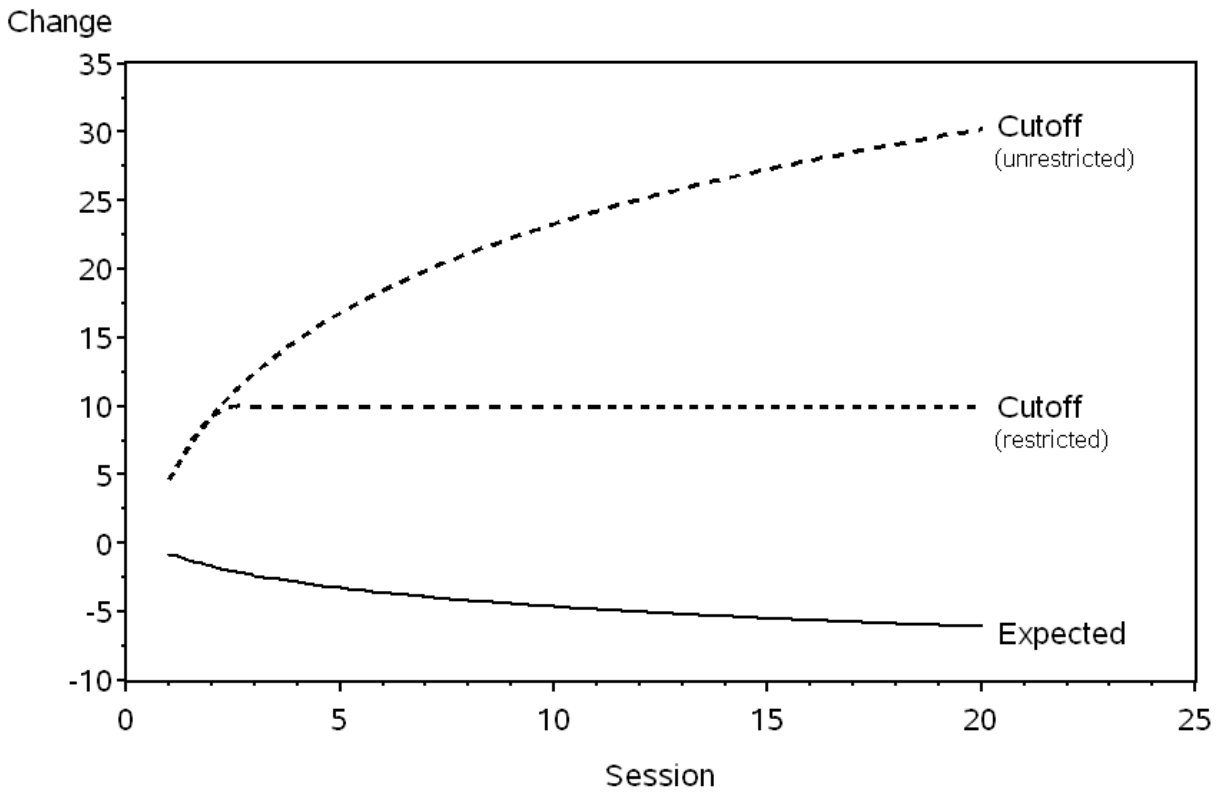

Figure 6. Modeled change scores and related cutoffs.

Warning System Prediction Accuracy

A primary purpose of the warning system was to identify clients at risk for deterioration. The primary purpose of the second part of this study was to test the accuracy of a system that made predictions of deterioration using the cutoffs described above. We used these cutoffs produced from the reference sample to predict which clients in the validation sample would have a final outcome of deterioration. We assigned clients in the validation sample the cutoff scores

corresponding to their baseline scores. If any observed score, other than the first or last, ever reached or surpassed the cutoffs, we signaled the client as having a predicted outcome of deterioration. We did not allow a first or last score to signal a client because these were the two scores used to identify the actual outcome. By separating the scores used for prediction from those used for determination of actual outcome, we were careful to avoid potentially inflating the accuracy of our predictions. However, given the frequency of unplanned termination of treatment in actual practice, clinicians may rarely know which score is the last. By definition, final scores for clients who deteriorate are elevated and may often reach the warning system cutoffs. A warning system that in practice uses these final scores for predictions would likely yield superior prediction accuracy compared to the system reported in this study. Further, in a system using our approach of restricting cutoffs to a change score of 10 points (corresponding to the YOQ's RCI value), final YOQ scores for actual deteriorators will by definition signal these clients as having reached the cutoffs and at risk for deterioration.

We classified clients signaled by the cutoffs as predicted positives for deterioration and the nonsignaled clients as the predicted negatives (i.e., deterioration vs. non-deterioration). Table 19 cross tabulates our predicted outcomes by row with the actual observed outcomes by column. For clients the warning system cutoffs predicted to deteriorate, the true positives are the clients who actually did deteriorate and the false positives are the clients who did not deteriorate. For clients the warning system cutoffs predicted to not deteriorate, the true negatives are the clients who actually did not deteriorate and the false negatives are the clients who did deteriorate. Streiner (2003) suggested that studies report such values to enable readers to perform their own calculations of prediction accuracy and to double-check the calculations presented in the study. We used the values of this table in our calculations of prediction accuracy that follow.

Table 19

Cross Tabulation of Predicted and Actual Outcomes

	Actual			
	Raw scores		Change scores	
Predicted	Deterioration	Non-deterioration	Deterioration	Non-deterioration
Deterioration	68 (TP)	166 (FP)	71 (TP)	129 (FP)
Non-deterioration	43 (FN)	599 (TN)	41 (FN)	635 (TN)

Note. TP = true positives, FP = false positives, TN = true negatives, FN = false negatives.

Table 20 presents the accuracy with which our warning system cutoffs predicted actual outcomes of deterioration versus non-deterioration. Other than the likelihood ratio, each value listed in the table can be understood as the percentage of clients of a certain type that the warning system identified with an early warning signal. The percentages are calculated as ratios (Streiner, 2003). For example, sensitivity is calculated as the number of deteriorating clients the system identified divided by the total number of deteriorating clients ——— . The sensitivity values listed in Table 20 indicate that the warning system's raw score cutoffs correctly identified 61% of the clients in the validation sample that actually deteriorated, versus 63% for the change score cutoffs.

The specificity values ——— in Table 20 indicate that the warning system's raw score cutoffs correctly identified 78% of the clients in the validation sample that did not deteriorate, versus 83% for the change score cutoffs. The hit rate values,

———— ,

in Table 20 indicate that the warning system's raw score cutoffs were correct in 76% of their classifications, versus 81% for the change score cutoffs. The likelihood ratio values,

———— ———— ,

in the table indicate that using the raw score cutoffs, a prediction to deteriorate was 2.82 times more likely for a client who actually deteriorated than for a client who did not, versus 3.78 times more likely using the change score cutoffs.

Table 20

Prediction Accuracies of Standard Warning System Cutoffs

Method	Sensitivity	Specificity	Hit rate	Likelihood ratio for deterioration	Positive predictive power	Negative predictive power	% of false positives that show no change
Raw score	.61	.78	.76	2.82	.29	.93	71%
Change score	.63	.83	.81	3.78	.35	.94	74%

Note. These prediction accuracies were calculated using subsample 2, for which *n* ranged from 874 to 879 in the 10 iterations of random samplings.

The values for positive predictive power ——— indicate that of all the clients predicted to deteriorate using the raw score cutoffs, 29% actually deteriorated, versus 35% for the change score cutoffs. These values are low likely due to deterioration comprising a relatively small percentage of the sample, a phenomenon discussed by Streiner (2003). The values for negative predictive power ——— indicate that of all the clients predicted to not deteriorate using the

raw score cutoffs, 93% actually did not deteriorate, versus 94% for the change score cutoffs. These values are high likely due to non-deterioration comprising a relatively large percentage of the sample, a phenomenon also discussed by Streiner. The positive predictive powers of 29% and 35% for the raw score cutoffs and the change score cutoffs imply that 71% and 65% of clients predicted to deteriorate did not.

The final column of Table 20 presents the percentages of the false positives whose outcome demonstrated no reliable change (i.e., final score was not reliably different from baseline, as per the RCI value requiring a minimum 10 point change). Although 71% of the clients that the raw score cutoffs predicted to deteriorate did not deteriorate, 71% of these false positives did not make any reliable improvement and could likely have benefited from the extra clinical attention nonetheless. Of the 65% of clients the change score cutoffs predicted to deteriorate but who did not, 74% did not make any reliable improvement.

Prediction accuracy of alternative cutoffs. Examining the different prediction accuracies between the raw score cutoffs and the change score cutoffs, we recognized the possibility that the slightly higher accuracy of the change score approach may have been due to its restriction of the cutoff at a change score of 10 points. We explored this potential phenomenon by calculating the accuracy of the raw score approach while applying a similar 10-point restriction on cutoffs' deviation from baseline. Conversely, we calculated the accuracy of the change score approach using cutoffs no longer restricted to a change score of 10 points, but extending higher (as illustrated in Figure 6). Table 21 presents the prediction accuracy of these and other alternative approaches to creating the warning system's cutoff scores. The prediction accuracy for the original raw score approach appears in Trial 1 on the table. Trial 2 presents the accuracy once a 10-point change score restriction was applied to the cutoffs of this raw score approach. The

sensitivity improved from .61 to .65, with a no change to the specificity or to the hit rate. The prediction accuracy for the original raw score approach appears in Trial 16 on the table. Trial 17 presents the accuracy once the 10-point change score restriction was removed from the cutoffs of this approach. All indices remained unchanged.

Table 21 categorizes the several variable options we explored in creating warning system cutoff scores. The second column indicates whether the YOQ scores being monitored were raw scores or change scores, alternatives that have been explored in detail above. The third column introduces a new option for whether the cutoff scores are generated using prediction intervals, as in all approaches discussed to this point, or whether they are base solely on a prescribed change score. The prediction interval basis allows for cutoffs that have a nonzero slope, whereas the change score basis is a flat line cutoff corresponding to a chosen change score. An example of the latter appears in Figure 7 and corresponds to Trial 22 of Table 21. The conceptual impetus for basing cutoffs on prediction intervals was to identify a selected percentage of clients whose scores were worsening relative to their baseline. The fourth main column in Table 21 specifies the chosen percentage when prediction intervals are used as the basis for creating the cutoff scores. Trials 1 and 16 show the original two approaches to creating the warning system cutoffs, each of which used prediction intervals to identify 14.5% of clients, corresponding to the deterioration rate of the reference sample. The fifth main column specifies the change score to which the cutoff was restricted, which could be applicable while monitoring raw scores or change scores and while the cutoffs are based on prediction intervals or simply on the change scores themselves.

Trials 3–13 show the results of experimenting with a series of increasing percentages of clients to be identified by cutoffs based on prediction intervals. As the identified percentage

Table 21

Prediction Accuracies of Alternative Warning System Cutoffs: A

Trial	Scores examined	Cutoff basis	Percentage to identify	Cutoff restriction (change score)	Sensitivity	Specificity	Hit rate	Likelihood ratio for deterioration	% of false positives that show no change
1	raw	pred	14.5%	—	.61	.78	.76	2.82	71%
2	raw	pred	14.5%	10	.65	.78	.76	2.93	72%
3	raw	pred	10.0%	10	.62	.82.	.79	3.36	72%
4	raw	pred	12.0%	10	.64	.80	.78	3.14	73%
5	raw	pred	14.0%	10	.65	.78	.76	2.98	72%
6	raw	pred	16.0%	10	.66	.76	.75	2.81	71%
7	raw	pred	18.0%	10	.68	.75	.74	2.73	71%
8	raw	pred	20.0%	10	.68	.74	.73	2.60	70%
9	raw	pred	22.0%	10	.70	.72	.72	2.53	70%
10	raw	pred	24.0%	10	.71	.71	.71	2.43	71%
11	raw	pred	26.0%	10	.72	.69	.69	2.34	70%
12	raw	pred	28.0%	10	.74	.68	.68	2.27	70%
13	raw	pred	30.0%	10	.75	.66	.67	2.20	70%
14	raw	pred	67.6%	—	.85	.34	.40	1.28	64%
15	raw	pred	67.6%	10	.85	.34	.40	1.28	64%
16	change	pred	14.5%	10	.63	.83	.81	3.78	74%
17	change	pred	14.5%	—	.63	.83	.81	3.78	74%
18	change	pred	14.5%	5	.70	.78	.77	3.19	76%
19	change	change	—	10	.53	.89	.85	4.90	77%
20	change	change	—	9	.57	.87	.83	4.51	76%
21	change	change	—	8	.61	.86	.83	4.35	76%
22	change	change	—	7	.63	.84	.81	3.87	76%
23	change	change	—	6	.67	.81	.79	3.54	76%
24	change	change	—	5	.68	.79	.77	3.20	76%
25	change	change	—	4	.69	.76	.75	2.90	77%
26	change	change	—	3	.73	.73	.73	2.64	77%
27	change	change	—	2	.73	.70	.70	2.40	77%

Note. These prediction accuracies were calculated using subsample 2, for which *n* ranged from 874 to 879 in the 10 iterations of random samplings. Pred = prediction interval as basis for creating cutoffs.

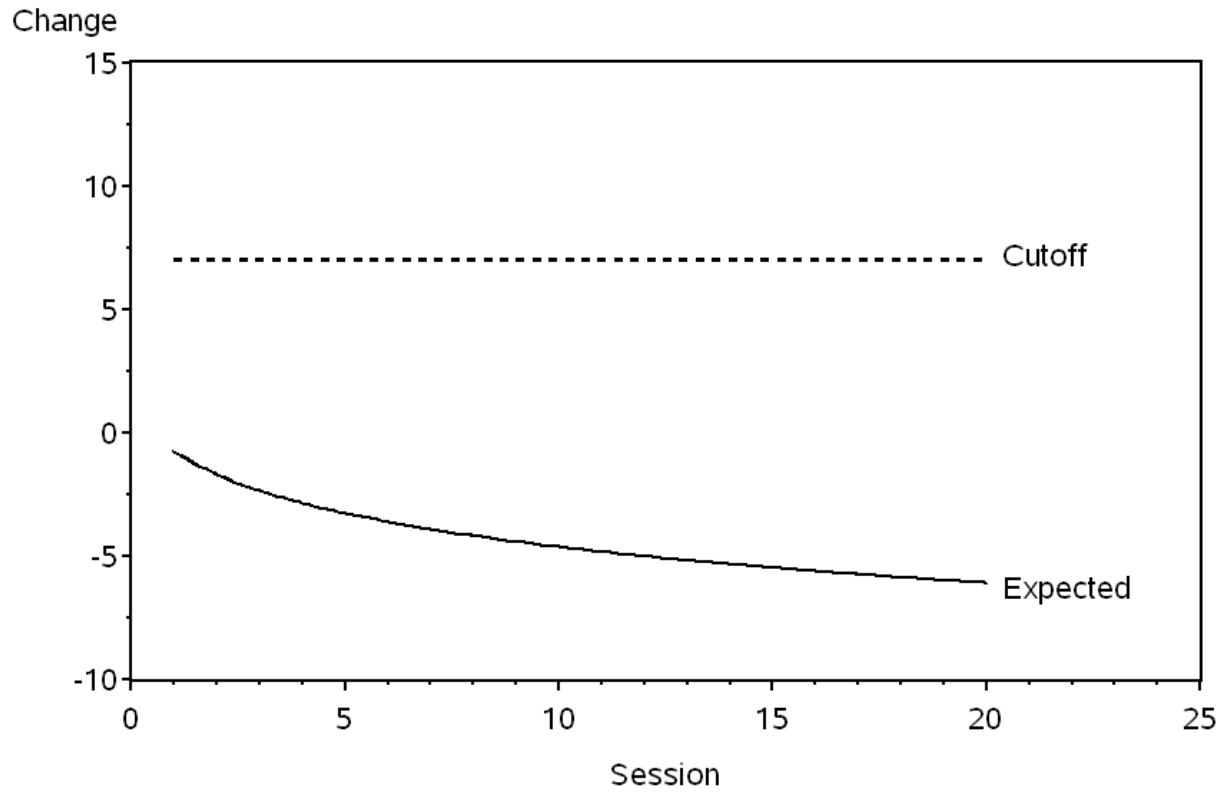

Figure 7. Modeled change scores with cutoff equal to a change score of 7.

increased, the sensitivity increased, but apparently at the expense of specificity and hit rate. Trials 14 and 15 use the percentage corresponding to the percentage of clients in the reference sample experiencing no reliable improvement (52.1% + 13.3% + 1.2% = 66.6%; see Table 16). This resulted in higher sensitivity, but substantially lower hit rate, similar to the trend for the increasing percentages identified in Trials 3–13. Trial 18 presents the results of restricting the prediction interval cutoffs based on change score to 5 points, which again boosted sensitivity at the expense of specificity and hit rate.

Trials 19–27 in Table 21 present the results of a series of cutoffs abandoning the prediction intervals altogether in favor of simply examining a predetermined change score as the

basis for the cutoffs. Again, the effect is for the cutoffs to simply be flat lines, as illustrated in Figure 7. Consistent to the pattern associated with using more stringent cutoffs, the successively diminishing change score restrictions improve sensitivity at the expense of specificity and hit rate. Trial 22 of Table 21 shows that a simple cutoff placed at a change score of seven points achieves prediction accuracies better than any variation on the prediction interval approaches reported in this study.

The above approaches to creating warning system cutoffs used either raw scores in separate models per score band, or change scores in a single model. Given that the purpose of cutoffs is only to identify YOQ score deviations (i.e., change scores) equal to or greater than the YOQ's RCI value of 10 points, the change score approach may be the broader or more general approach. The raw score approach pursues more specificity in that it requires some kind of accommodation for varying baseline scores. The common approach of creating score bands to account for varying baseline scores has the tradeoff of limited sample sizes per model per score band. An alternative that could account for baseline scores—while still modeling raw scores rather than change scores—could be to include some kind of predictor variable in the model that accounts for baseline. The predictor would have the effect of shifting the prediction intervals higher or lower to accommodate each client's baseline score or trajectory elevation.

Table 22 presents the comparative prediction accuracies of these alternative methods of accounting for baseline score. Trial 1 presents the prediction accuracies for the original raw score approach, this time with the 10-point change score restriction on the cutoffs. Trial 2 presents the prediction accuracies for the original change score approach. Trial 3 presents the recently proposed alternative to the original score band approach by modeling the entire sample (rather than separate score bands) and including a predictor variable for score band (centered around its

mean). This alternative yielded almost equal accuracy, with only a slightly lower sensitivity (.63 vs. .65) yet a slightly higher hit rate (.78 vs. .76). Trial 5 presents the accuracy of a slight variation on this alternative by substituting the quasi categorical–continuous variable for score band with the continuous variable for baseline score (centered around its mean). The resulting accuracy is superior to the approach using a variable for score band (Trial 3) and superior to the original approach of separate models per score band (Trial 1).

The prediction accuracies presented in Table 22 also offer the opportunity to demonstrate the inutility of anything but global cutoffs to predict the global phenomenon of YOQ scores changing by the RCI-based value of 10 points. Although intuition may tend toward anticipating that additional predictor variables in the multilevel models would yield cutoffs demonstrating superior prediction accuracy, such is not the case. Other than the main effect for clients' baseline scores, predictor variables only cause the modeled trajectory, and its corresponding prediction intervals or cutoffs, to deviate from the overall sample average. The resulting cutoffs would therefore be designed to signal as at risk for deterioration the global percentage of a nonglobal group. A natural remedy could be to set the cutoffs to correspond to the deterioration rate for a particular subgroup, but little would be gained because deterioration still has the global definition of 10 or more points worsening. The specialized cutoffs would only be working to predict which clients would end up with 10 or more points worsening, which would be the same effect of the global cutoffs. The global and specialized cutoffs would be distinct only in their origins from different deterioration rates; their actual cutoffs for clients with equal baselines would be roughly equal. This specialized avenue of arriving at roughly equal cutoffs may be unnecessarily complicated, if not less favorable due to the smaller subsamples upon which it would have to rely for determination of deterioration rates and modeling prediction intervals.

Table 22

Prediction Accuracies of Alternative Warning System Cutoffs: B

Trial	Scores examined	Separate models per score band	Fixed effects	Sensitivity	Specificity	Hit rate	Likelihood ratio for deterioration	% of false positives that show no change
1	raw	yes	*LNSESS*	.65	.78	.76	2.93	72%
2	change	no	*LNSESS*	.63	.83	.81	3.78	74%
3	raw	no	*LNSESS*, score band	.63	.80	.78	3.19	73%
4	raw	no	*LNSESS,* score band, *LNSESS* *score band	.64	.78	.76	2.92	72%
5	raw	no	*LNSESS,* baseline	.65	.81	.79	3.38	73%
6	raw	no	*LNSESS,* baseline, *LNSESS* *baseline	.66	.77	.76	2.92	74%
7	change	no	*LNSESS,* baseline	.66	.80	.78	3.26	74%
8	change	no	*LNSESS,* baseline, *LNSESS* *baseline	.66	.76	.75	2.78	71%

Note. All using prediction intervals aiming to identify 14.5% in validation sample, and with cutoff restriction of 10-point change scores. These prediction accuracies were calculated using subsample 2, for which *n* ranged from 874 to 879 in the 10 iterations of random samplings.

Table 22 demonstrates the lack of increased accuracy when additional predictor variables are added to the model. Trials 4, 6, and 8 differ from the trials immediately preceding each in terms of the addition of an interaction of a baseline-related variable with the time variable *LNSESS*, having the effect of accounting for the differing rates of change according to baseline scores. This interaction term in the model likely accounts for more variability in slopes than any other. Comparing the accuracies of Trial 4 to Trial 3, Trial 6 to Trial 5, and Trial 8 to Trial 7, it is clear that in this instance, the addition of a strong predictor variable did not create warning system cutoffs any better than if the predictor variable had been omitted.

Table 22 demonstrates that various approaches to creating warning system cutoffs in terms of raw scores, each having to account for variability in initial score, do not yield prediction accuracies at all superior to cutoffs based on change scores. To reiterate, this is likely because the purpose of the cutoffs is so tied up in predicting the global RCI-based change score of 10 points or more. Accounting for anything other than change score may add unnecessary noise to the procedure. Note, too, the argument likely true to the patient-focused research paradigm that outcome predictions should rely on what the clinician sees in a specific client's ongoing outcomes (change scores), as opposed to making predictions according to predetermined generalizations associated with this client's demographic (as one might attempt to account for by including additional predictor variables).

Incorrect predictions. We hypothesized that our incorrect predictions of deterioration or non-deterioration would be associated with particular YOQ trajectory shapes. Our predictions were correct for clients whose trajectories steadily inclined toward deterioration or declined toward recovery. Our predictions were incorrect for clients whose trajectories showed a change in directionality. In Figure 8 we summarized trajectory shapes for true positives, false positives,

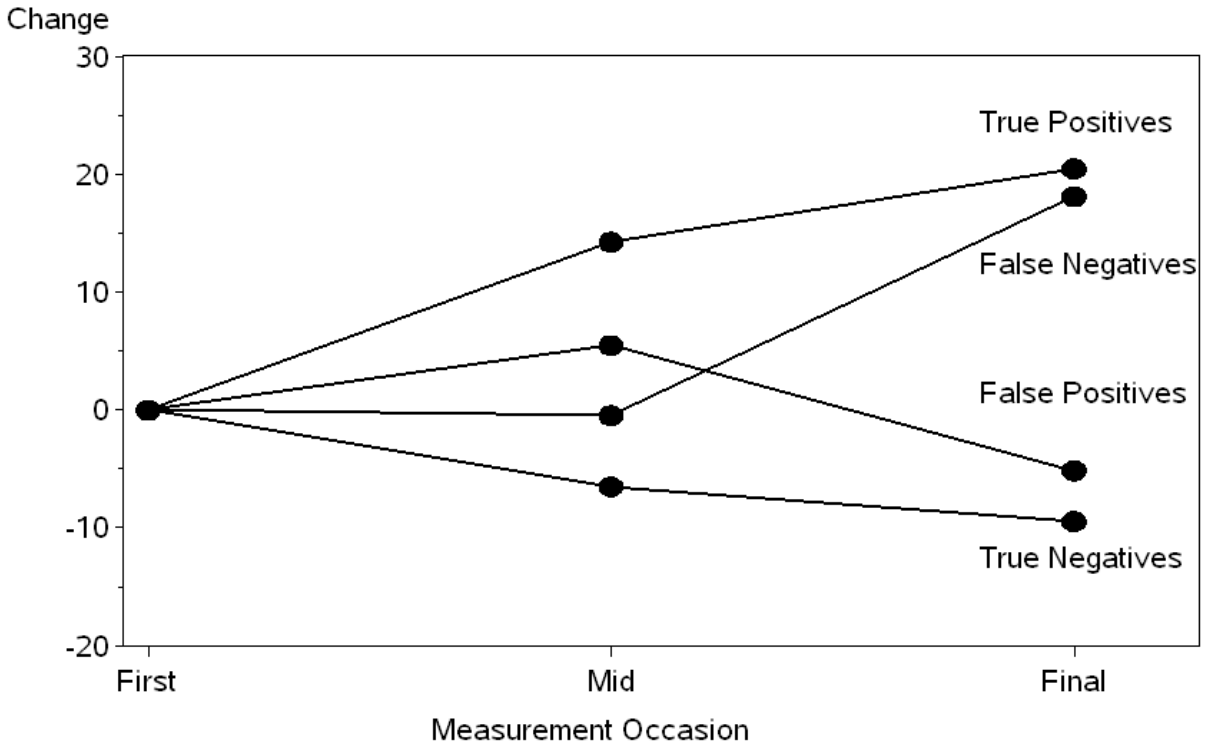

Figure 8. Trajectory shapes for clients predicted correctly and incorrectly for deterioration using cutoffs based on raw scores. All midtreatment change scores collapsed into a single mean change score.

true negatives, and false negatives. Each trajectory summary consists of three data points. The first and last data points correspond to clients' first and last YOQ change scores. The middle data point corresponds to the mean of the midtreatment change scores (scores that are neither the first nor the last). A line connecting the first and second data points depicts a general trajectory direction, which may or may not continue in approaching the final data point. Figures 8 and 9 illustrate how our predictions were often incorrect for clients whose trajectory shapes included a change in general direction. Further exploration of relationship between trajectory shape and

prediction accuracy is beyond the scope of this study, but could potentially play a role in the development of improved warning system approaches to identifying clients at risk for negative outcome.

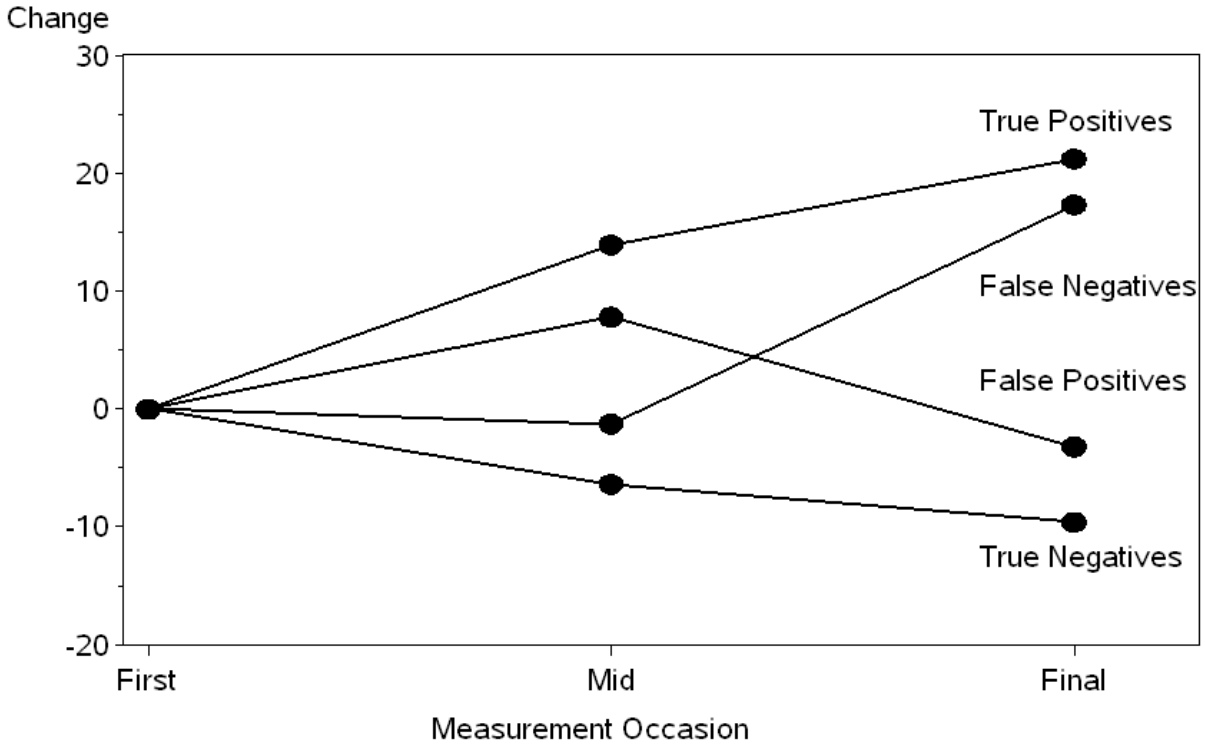

Figure 9. Trajectory shapes for clients predicted correctly and incorrectly for deterioration using cutoffs based on change scores. All midtreatment change scores collapsed into a single mean change score.

DISCUSSION

The field of mental health treatment is making efforts to better serve all psychotherapy clients, especially clients such as the 13% of youth in this study's larger sample who experienced a significant worsening of scores on the Youth Outcome Questionnaire-30 (YOQ; Burlingame et al., 2004), or the broader 71% who did not experience a reliable improvement. These efforts to improve psychotherapy services involve collaboration between research and practice because therapists on their own are less accurate in predicting which clients will experience negative outcome. The patient-focused research paradigm has shifted the field's focus from group-level treatment outcomes to outcomes on the individual client level, including outcome monitoring for purposes of treatment planning and quality care. Some of these monitoring systems include early warning systems to help identify and better serve clients who are at risk for negative outcome.

Summary and Implications

Part 1 of the present study validated previous studies by identifying variables that were predictive of youth change trajectories on the YOQ. Part 2 of this study replicated tests of the accuracy of a warning system for at-risk youth clients, using the YOQ. This process compared various approaches to creating the cutoffs the warning system used to make its predictions of clients' final outcome. These cutoffs achieved prediction accuracies that appear to warrant the next step of testing whether the application of such a warning system for youth in clinical practice clients yields improved outcomes, as has been demonstrated for similar warning systems used with adult clients.

YOQ Change Trajectories

In Part 1 of this study we created multilevel models of YOQ scores over time to identify the portions of variability attributable to clients versus therapists and to identify other relevant

predictor variables. For our sample, it appeared that 7% of the overall variability in YOQ scores was attributable to therapists, which appears near or slightly more than that found in similar studies (Cannon et al., 2010; Wampold & Brown, 2005; Warren, Nelson, & Burlingame, 2009; Warren et al., 2010). These similar studies found a small portion of the variability to be attributable to treatment site as well, but such effects were nonsignificant in the present study, likely due to the vast majority of services being provided at a single site. We noted that 8% of variability in trajectory elevations was related to differences between therapists. Somewhat more notable, however, was our finding that 16% of variability in trajectory slopes was associated with differences in therapists. Incidentally, Wampold and Brown found roughly 5% of variability in scores on an adult version of the YOQ (i.e., the Outcome Questionnaire, Lambert et al., 2004) to be associated with differences in therapists, and drew on data from the same managed care setting from which data were obtained for the present study. The higher percentages of variability attributable to therapist in the current study may possibly be associated with greater variability in levels of experience and training that therapists have in working with the youth.

To identify variables predictive of YOQ scores over time, we created a multilevel model with a number of hypothesized predictors. Not all were significant in the model, so we used a number of iterations of model building (including but not limited to stepwise deletion and stepwise addition of variables) to arrive at a final model of variables predicting YOQ scores over time. This final model is best illustrated in Figure 4. The figure demonstrates that the predictor variable likely of the most clinical significance are the following: Clients with recent treatment—that is, their current outpatient treatment episode began within 90 days of treatment in the inpatient or day treatment settings—had a trajectory elevation roughly eight points higher than that of other clients. Yet, this variable or characteristic is not associated with differences in YOQ

rate of change. On the other hand, clients with prior treatment that was not so recent—more than 90 days prior to the current treatment episode—had a trajectory elevation only 2.3 points higher, but a substantially slower rate of change than average.

Clients with more YOQs per month appeared to have slightly lower baseline scores and faster rates of change. We considered the possibility that this could be merely an artifact of the clients with more frequent YOQs simply being those who terminated treatment in the early stages during which YOQs were administered more frequently. However, the effect persisted even when our model controlled for the effects of episode duration simultaneously in terms of total weeks and total months for the current treatment episode. Future studies could explore what might account for the relationship between more frequent measurement and faster rates of change. One likely explanation is that more frequent measurement and feedback to clinicians is associated with improved outcomes for youth clients. This is very encouraging for the general aims of this study, suggesting that an early warning system that provides clinicians this feedback may be rather beneficial with youth, as has been demonstrated with adults.

Warning System Cutoffs and Accuracy

A common implementation of the warning system proposed in this study is for clinicians to be alerted to clients whose scores reach or surpass the cutoffs. Clinicians may use their judgment as to what additional attention will be appropriate for each given client, but one approach would be to administer additional measures exploring factors often associated with psychotherapy outcomes (e.g., therapeutic alliance, motivation to change, social support network, etc.). Clients whose therapists received feedback from such a system have experienced improved outcomes (Harmon et al., 2007; Hawkins et al., 2004; Lambert, Whipple, et al., 2001; Lambert et al., 2002; Whipple et al., 2003). Compared to at-risk clients in the nonfeedback

condition, nearly twice as many at-risk clients from the feedback condition ended treatment with improvement (9 clients vs. 4) and even more ended with recovery (i.e., final scores in the nonclinical range; 5 clients vs. 1). These superior outcomes may be due to the at-risk clients in the feedback condition receiving twice as many sessions on average (9.3 sessions vs. 4.7), presumably as a result of the feedback. In addition, it appears that simultaneous feedback to therapists and their clients may achieve even better outcomes than when only therapists receive feedback (Hawkins et al., 2004).

To be clear, the warning system this study proposed is not intended for use in assessing the effectiveness of particular therapists or treatment modalities. Rather, it is designed as an idiographic assessment of client outcomes in a single context. Its purpose is to provide clinicians added data to evaluate using their clinical judgment. This raises a crucial issue. Although this and past studies have demonstrated adequate prediction accuracies associated with warning systems such as this, the warning system's success and utility nonetheless is completely vulnerable to whether clinicians have sufficient instruction and motivation to use the system. At the extreme, the mere mention of outcome classes could be met with defensiveness from clinicians invested in their clients' outcomes and their own therapeutic effectiveness.

The most central purpose of this study was to test and demonstrate the potential accuracy an early warning system could have in predicting which clients were at risk for negative outcome in terms of a significant increase in YOQ scores. Similar to past studies, we designed the system to make its predictions based on cutoffs against which clients' observed scores would be compared over the course of treatment. We tested the accuracy of cutoffs created using two different approaches. Our evaluation of these approaches inspired our testing of a series of alternative approaches to creating potential warning system cutoffs, but also to distinguish the

meaningful considerations in creating these cutoffs from the considerations that appear unnecessary. Finally, we identified YOQ trajectory shapes associated with clients for whom our outcome predictions were incorrect.

We based our two primary approaches to creating warning system cutoffs based on the upper boundary of a *t*-type confidence interval created around YOQ scores modeled using multilevel modeling. As shown in Table 18, clinicians could use these cutoffs to identify clients at risk for deterioration. Figures 5 and 6 provide a visual illustration of how the cutoffs compare to the expected YOQ scores. We created the cutoffs using a reference sample and then tested their predictive accuracy in a validation sample. Similar to past studies' warning system cutoffs based on raw scores, our cutoffs based on raw scores produced predictions of deterioration achieving a sensitivity of .61, a specificity of .78, and a hit rate of .76. Our cutoffs based on YOQ change scores produced predictions of deterioration achieving only slightly higher accuracy, with a sensitivity of .63, a specificity of .83, and a hit rate of .81. The hit rates of these two approaches are consistent with similar past studies, whose hit rates ranged from .69 to .88 (Bishop et al., 2005; Bybee et al., 2007; Cannon et al., 2010; Lambert et al., 2002; Warren et al., 2009). Sensitivities from these past studies were somewhat higher than the present study, ranging from .61 to .77.

It is likely that the warning system tested in this study would achieve higher prediction accuracies in actual practice. The accuracies we reported stem from our conservative approach of omitting final YOQ scores from those we used to predict final outcome. Clinicians using such a warning system would be using all YOQ scores for prediction (other than the baseline), including the final YOQ score. This final score is typically high for clients with negative outcomes and would likely alert clinicians to give these clients extra attention. Further, in a

system using our approach of restricting cutoffs to a change score of 10 points (corresponding to the YOQ's RCI value), final YOQ scores for actual deteriorators would by definition signal these clients as having reached the cutoffs and as at risk for deterioration.

Characteristics of optimal cutoffs. Our evaluation of the prediction accuracies of the above warning system cutoffs based on raw scores and based on change scores led us to identify several important considerations in creating these cutoffs. The first consideration was whether the warning system would compare its cutoffs to raw scores or change scores from the YOQ. Change scores may be the simpler broader case, whereas raw scores may introduce complexities that have intuitive appeal, but extend beyond the very basic and limited nature of the RCI-based definition of deterioration and outcome classes. This study's various approaches to creating warning system cutoffs in terms of raw scores, each having to account for variability in initial score, did not yield superior prediction accuracies compared to cutoffs based on change scores. This is likely because the purpose of the cutoffs was almost exclusively to predict the global RCI-based change score of 10 points or more. Accounting for anything other than change score may add unnecessary complexity to the procedure, which may account for this study's slightly higher prediction accuracy associated with cutoffs based on change scores compared to raw scores.

The second consideration for creating the warning system's cutoffs was whether they would be based on prediction intervals or simply based on change scores. Cutoffs based on prediction intervals aim to identify predetermined percentages of the most severe YOQ scores and facilitate cutoffs that change over time. In contrast, cutoffs based on change scores are simply flat, always equal to a predetermined deviation from the baseline YOQ score. The results

of this study demonstrated that with appropriate specifications, both approaches yielded roughly equal prediction accuracies.

The third consideration for creating the warning system's cutoffs was whether to restrict the sloping cutoffs (based on prediction intervals) to a predetermined maximum change score. A cutoff restriction of 10 points—corresponding to the YOQ's RCI value indicating the minimum amount of change that can be considered distinguishable from measurement error—occasionally improved the prediction accuracies and did not ever appear to diminish them. Future studies may check whether the benefits of such cutoff restrictions are consistent with other data.

The fourth consideration for creating the warning system's cutoffs was whether to include prediction variables in the unconditional growth model, which included only a predictor variable for time in order to account for slope. Similar to the intuitive appeal of examining raw scores over change scores, a common expectation could be that additional predictors in the model would customize the resulting cutoffs and thus increase the prediction accuracy. With no predictor variables in the model, the cutoffs are created by a very global means; they correspond to the upper boundary of a prediction interval for the unconditional growth model. This is a global means toward the global end of identifying clients who will have an overall worsening change score of 10 points or more, the definition of deterioration for the YOQ. The addition of predictor variables may inappropriately create a customized or specific means to the same global end. Until the end is customized (e.g., RCI values or definitions of deterioration specific to subpopulations) and no longer global, the added complexity may have no apparent benefit.

Supporting the conceptual argument above, the results in this study demonstrated no added value to prediction accuracy when warning system cutoffs came from prediction intervals whose models included extra predictor variables other than a time variable to account for slope,

and possibly a variable to represent the baseline if raw scores were being used. The omission of any demographic predictor variables in favor of only monitoring observed outcomes in relation to global cutoff scores may demonstrate some conceptual consistency with the aims of the patient-focused research paradigm; the outcome predictions rely on what a clinician actually observes in a specific client's ongoing outcomes, as opposed to making predictions according to generalizations associated with the client's demographic.

In summary of these considerations, results from the present study suggest that the best practices in creating warning system cutoffs may be as follows. Warning system cutoffs may be equally effective whether simply a change score shown to be appropriate for or generalizable to the population at hand, or cutoffs based on prediction intervals associated with multilevel models of scores over time. If the prediction interval approach is taken to creating cutoffs, it may be simplest and most accurate if modeling change scores rather than raw scores and if it includes no predictor variables other than a time variable and possibly a variable accounting for variability in baseline scores. Finally, if the cutoffs are based prediction intervals, they may yield slightly higher prediction accuracy if restricted to a maximum change score corresponding to the measure's RCI value.

Inaccurate predictions. False positives are often a concern in screening or warning systems, sometimes with costly or dangerous consequences. In the case of the present study, note that although 71% of the clients that the raw score cutoffs predicted to deteriorate did not deteriorate, 71% of these false positives did not make any reliable improvement and could likely have benefited from the extra clinical attention nonetheless. Similarly, of the 65% of clients the change score cutoffs predicted to deteriorate but who did not, 74% did not make any reliable improvement. It appeared that the majority of false positives associated with this study's warning

system were not progressing in treatment as would be hoped and could likely have benefitted from the added clinical attention.

We compared YOQ score trajectories for clients for whom our outcome predictions with trajectories for clients for whom our predictions were incorrect. As shown in Figures 8 and 9, our predictions were correct for clients whose trajectories followed a consistent trend upward or downward. In contrast, our cutoffs most commonly yielded incorrect predictions for clients whose trajectories trended upward, reaching the cutoffs to signal the clients as predicted to deteriorate, yet having a lower final YOQ score. These clients constituted the false positives for deterioration. Clients who were false negatives most commonly had trajectories that trended downward, apparently progressing appropriately in treatment, yet having a sufficiently high final YOQ score to constitute deterioration. It is notable that in our attempt to be conservative, we did not include in our calculations of prediction accuracy the warning signals that would have been generated or nullified by these final YOQ measurements. Actual clinical application of the warning system would benefit from examining these final scores, thus avoiding the majority of the false positives and false negatives reported in this study.

Limitations

The administration frequency of the YOQ was relatively good (at sessions 1, 3, 5, 10, 15, 20, etc.) and demonstrated that one managed care organization found it feasible to administer an outcome measure as part of routine services. However, ideal data would have included YOQ administrations at each session, facilitating more accurate and reliable measurement, but possibly greater opportunity for false positives with the warning system. Given our constraint of requiring two YOQs per client in the Part 1 analyses and 3 YOQs per client in the Part 2 analyses, YOQ administration at each session would have allowed clients with shorter treatment episodes (in

terms of sessions) to have been included. Inclusion criteria related to YOQ administration was responsible for the greatest amount of archival data we disqualified from inclusion in this study. Our samples represented only 31% and 13% of the original archive, for the analyses of parts 1 and 2 respectively. Small percentages such as these admittedly may not reflect the larger archive. However, the sample selected for the calculation of prediction accuracies has very similar characteristics to the subpopulation to which its results are intended to generalize. The warning system is primarily only useful for clients having the characteristics corresponding to our selection criteria, especially in terms of numbers of YOQ measurements.

The aforementioned issue of generalizability is important and comes into play considering the split samples approach we used to create and test the accuracy of the warning system's predictions of outcome. We took care to assign clients to the reference and validation samples at random in order to avoid systematic differences between the samples that could serve as confounds and artificially inflate or deflate prediction accuracies. We also repeated the random assignment process ten times, reporting the mean results of the ten iterations of analyses with different random samplings. Nonetheless, all client data was produced in the same handful of clinical locations. The particular deterioration rate and warning system cutoffs we created in this study may not be fully generalizable to other differing clinical settings. We acknowledge that our data came from an outpatient managed care facility serving youth of average to above-average socioeconomic status. We offer the caveat that we do not intend this study's specific deterioration rate and warning system cutoffs to be applied in other settings. Instead, we intend this study to be a proof of concept, that a warning system can be created and applied specifically for a particular clinical location's deterioration rate and other characteristics.

More generally, a larger sample size could have enhanced this study. More clients would have facilitated the creation of a greater number of score bands, each with narrower baseline ranges, as used in one raw score approach to creating warning system cutoffs. It would have been helpful had the data included information regarding clients' race. Our results would likely have been different if we had also selected self-report YOQ measurements from the archive. Cannon et al. (2010) tested the comparative prediction accuracies of warning systems accounting for self-report and parent-report YOQs, the combination of the two yielding the highest prediction accuracy. In addition, this study included the YOQ as its only outcome measure. Although the YOQ is designed to be a broad measure of global functioning, the lack of other outcome measures may have limited this study's perspective on outcome. On a related note, some readers may disagree with deterioration in treatment being defined as a worsening of 10 or more points on the YOQ, taking issue with the single measure, or perhaps with the notion of the reliable change index of 10 points being global and insensitive to any particular demographic. However, the approach of having a single outcome measure may be a key characteristic of an outcome monitoring system that remains feasible in clinical practice.

An additional limitation may be the unknown yet possible ways in which a warning system for youth may differ from a warning system for adults. Application of a warning system for youth is not as widely tested as for adults. In addition, deterioration, or premature termination of treatment, may have added complexities for youth. Youth are likely more susceptible than adults to external factors (e.g., parent and family considerations) affecting to their therapy outcomes and therapy attendance. Psychotherapy for youth commonly includes other complications beyond those typical for psychotherapy for adults, one example being therapists serving youth without the appropriate training.

Future Directions

This study further examined predictors of psychotherapy outcome in terms of YOQ trajectories and demonstrated the potential accuracy of an early warning system that could help clinicians give needed extra attention to the 71% of clients who simply do not show any reliable improvement in terms of YOQ scores. The most important next step in this line of research would likely be to test the results of implementing a warning system such as this in clinical practice. Similar warning systems for adult clients have helped improve psychotherapy outcomes and likely have the potential to do the same for youth. Further exploration of the underlying causes of deterioration may help uncover important aspects of helpful interventions for clients who do not appear to be benefitting from treatment.

Similarly, future studies could examine each of the predictor variables found to be associated with the elevation and slope of YOQ score trajectories as explored in Part 1 of this study, attempting to better understand the relationship between these variables and YOQ scores. It could be particularly important for studies to test whether YOQ measurement frequency is associated with improved outcomes in other data and settings; this appears to have bearing on the utility of outcome monitoring and even the implementation of an early warning system. Future studies could examine whether the procedures this study found most successful in creating accurate warning system cutoffs are equally important to the accuracy of warning system cutoffs created using different data. It would be appropriate to replicate these procedures using data from various types of treatment setting and from various respondents (e.g., self-report vs. parent-report YOQs). Other studies could also further explore the additional capacities the warning system has to identify not only deteriorators, but non-improvers also, a major portion of this study's sample that also should receive added clinical attention.

Finally, studies could explore the many issues that may be unique to monitoring outcomes in youth as opposed to adults. For example, increased outcome monitoring in youth often means more input from parents, which input may be especially helpful to clinicians whose youth clients are developmentally not as insightful or articulate. In addition, the present study demonstrated that a large portion of the variability in YOQ rate of change was associated with differences in therapists, which may underscore the greater variety in familiarity and training that clinicians have in working with youth versus working with adults. Another issue that may merit further exploration would be the relative lack of control youth have on their environment, and thus their psychotherapy outcomes, as compared with adults. Psychotherapy research for youth generally lags behind research for adults. This study and future studies can serve an important role in improving psychotherapy services for youth.

REFERENCES

Achenbach, T. M. (1991). *Manual for the child behavioral checklist/4–18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Anderson, E. M., & Lambert, M. J. (2001). A survival analysis of clinically significant change in outpatient psychotherapy. *Journal of Clinical Psychology, 57*, 875–888.

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

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., et al. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184–196.

Berrett, K. M. S. (1999). Youth Outcome Questionnaire: Item sensitivity to change. (Doctoral Dissertation, Brigham Young University, 1999/2000). *Dissertation Abstracts International, 60*, 4876.

Bishop, M. J., Bybee, T. S., Lambert, M. J., Burlingame, G. M., Wells, M G., & Poppleton, L. E. (2005). Accuracy of a rationally derived method for identifying treatment failure in children and adolescents. *Journal of Child and Family Studies, 14*, 207–222.

Bloom, A. (1987). Liability concern of utilization review and quality assurance programs. *HMO, 1*, 128–133.

Bobbit, B. L., Marques, C. C., & Trout, D. L. (1998). Managed behavioral health care: Current status, recent trends, and the role of psychology. *Clinical Psychology: Science and Practice, 5*, 53–66.

Brokowsky, A. (1991). Current mental health care environments: Why managed care is necessary. *Professional Psychology: Research and Practice, 22*, 6–14.

Brown, G. S., Lambert, M. J., Jones, E. R., & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care, 11*, 513–520.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Burlingame, G. M., Cox, J. C., Wells, M. G., Lambert, M. J., Latkowski, M., & Ferre, R. (2005). *The administration and scoring manual of the Youth Outcome Questionnaire*. Salt Lake City, UT: American Professional Credentialing Services.

Burlingame, G. M., Dunn, T., Cox, J., Wells, G., Lambert, M. J., & Brown, G. S. (2004). *Administration and scoring manual for the Youth Outcome Questionnaire-30 (YOQ-30).* Salt Lake City, UT: OQmeasures.

Burns, B. J., Hoagwood, K., & Mrazek, P. J. (1999). Effective treatment for mental disorders in children and adolescents. *Clinical Child and Family Psychology Review, 2*, 199–254.

Bybee, T. S., Lambert, M. J., & Eggett, D. (2007). Curves of expected recovery and their predictive validity for identifying treatment failure. *Dutch Journal of Psychotherapy, 33*, 419–434.

Canen, E. L., & Lambert, M. J. (May, 1999). *The incidence of patterned deterioration before stable improvement in psychotherapy.* Poster presented at the Western Psychological Association, Irvine, California.

Cannon, J. A. N., Warren, J. S., Nelson, P. L., & Burlingame, G. M. (2010). Change trajectories for the Youth Outcome Questionnaire Self-Report: Identifying youth at risk for treatment failure. *Journal of Clinical Child & Adolescent Psychology, 39,* 289–301.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388–400.

Cattani-Thompson, K. (2003). *The development of recovery curves for the Life Status Questionnaire as a means of identifying patients at risk for psychotherapy treatment failure.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.

Claiborn, C. D. & Goodyear, R. K. (2005). Feedback in psychotherapy. *Journal of Clinical Psychology: In Session, 61*, 209–217.

Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (sections 7.2, 7.8, and 9.2). Mahwah, NJ: L. Erlbaum Associates.

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

Davis, D. Thompson, M. A., Oxman, A. D., & Haynes. B. (1995). Changing physician performance: A systematic review of the effect of continuing medical education strategies. *Journal of the American Medical Association, 274*, 700–705.

Dawes, R. M. (1989). Experience and validity of clinical judgment: The illusory correlation. *Behavioral Sciences and the Law*, *7*, 457–467.

Docherty, J. P. (1999). Cost of treating mental illness from a managed care perspective. *Journal of Clinical Psychiatry, 60*, 49–53.

Donabedian, A. (1982). *The criteria and standards of quality.* Ann Arbor, MI: Health Administration Press.

Durlak, J. A., & McGlinchey, K. A. (1999). Child therapy outcome research: Current status and some future priorities. In S. W. Russ & T. H. Ollendick (Eds.), *Handbook of psychotherapies with children and families.* New York: Kluwer Academic/Plenum Publishers.

Fisher, D., Beutler, L. E., & Williams, O. B. (1999). STS clinician rating form: Patient assessment and treatment planning. *Journal of Clinical Psychology, 55*, 825–842.

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 and Psychotherapy, 8*, 231–242.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy.* Baltimore: Johns Hopkins University Press.

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analysis. *Psychological Bulletin*, *120*, 140–153.

Garland, A. F., Hurlburt, M. S., & Hawley, K. M. (2006). Examining psychotherapy processes in a services research context. *Clinical Psychology: Science and Practice*, *13*, 30–46.

Goldfried, M. R., & Wolfe, B. E. (1998). Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology, 66*, 143–150.

Grissom, R. J. (1996). The magical number .7+– .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology, 64*, 973–982.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.

Gunn, S. W. (1998). The quality imperative: An answer to the chaos of the behavioral health care environment. *Residential Treatment for Children and Youth, 16*, 35–65.

Haas, E., Hill, R., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology, 58*, 1157–1172.

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

Hansen, N. H. (1999). *An overview of longitudinal data analysis methodologies applied to the dose response relationship in psychotherapy outcome research.* Unpublished doctoral dissertation, Brigham Young University, Provo, Utah.

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.

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K., & Tuttle, K. (2004). The therapeutic effects of providing client progress information to patients and therapists. *Psychotherapy Research, 10,* 308–327.

Hoag, M. J., & Burlingame, G. M. (1997). Evaluating the effectiveness of child and adolescent group treatment: A meta-analytic review. *Journal of Clinical Child Psychology, 26*, 234–246.

Howard, K. I., Brill, P. L., Lueger, R. J., O'Mahoney, M. T., & Grissom, G. R. (1995). *Integra outpatient tracking assessment.* Philadelphia: Compass Information Services, Inc.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41,* 159–164.

Howard, K. I., Krause, M. S., & Lyons, J. S. (1993). When clinical trials fail: A guide to disaggregation. In L. S. Onken, J. D. Blaine & J. J. Boren (Eds.), *Behavioral treatments for drug abuse and dependence* (NIDA Research Monograph No. 137, pp. 291–302). Washington, DC: National Institute for Drug Abuse.

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

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

Howard, K., I., Orlinsky, D. E., & Lueger, R. J. (1995). The design of clinically relevant outcome research: Some considerations and an example. In M. Aveline & D. A. Shapiro (Eds.), *Research foundations for psychotherapy practice* (pp. 3–47). Sussex, England: Wiley.

Huffman, L. C., Martin, J., Botcheva, L., Williams, S. E., & Dyer-Friedman, J. (2004). Practitioners' attitudes toward the use of treatment progress and outcomes data in child mental health services. *Evaluation and the Health Professions, 27*(2), 165–188.

Ilardi, S.S., & Craighead, W.E. (1999). Rapid early treatment response, cognitive modification, and nonspecific factors in cognitive-behavior therapy: A reply to Tang and DeRubeis. *Clinical Psychology: Science and Practice, 6,* 295–299.

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.

Johnson, L. D., & Shaha, S. (1996). Improving quality in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 33*, 225–236.

Kadera, S., Lambert, M. J., & Andrews, A. (1996). How much therapy is really enough? A session-by-session analysis of the psychotherapy dose–effect relationship. *Journal of Psychotherapy Practice and Research, 5*, 132–151.

Kazdin, A. E. (1996). Dropping out of child therapy: Issues for research and implications for practice. *Clinical Child Psychology and Psychiatry, 1*, 133–156.

Kazdin, A. E. (2000). *Psychotherapy for children and adolescents: Directions for research and practice*. New York, NY: Oxford University Press.

Kazdin, A. E. (2003). Psychotherapy for children and adolescents. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change,* (5th ed., pp. 543–589). New York: John Wiley.

Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child and Adolescent Psychology, 34*, 548–558.

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

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology, 58,* 729–740.

Kopta, S. M., Howard, K. I., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology, 62*, 1009–1016.

Kordy, H., Hannover, 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. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143–189). New York: John Wiley and Sons.

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., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., et al. (2004). *Administration and scoring manual for the Outcome Questionnaire–45.* Orem, UT: American Professional Credentialing Services.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193). New York: Wiley.

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

Lambert, M. J., Whipple, J. L., 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*(1), 49–68.

Laurenceau, J. P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. *Clinical Psychology Review, 27*(6), 682–695.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science, 2*, 53–70.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

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

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

Lutz, W., Martinovich, Z., Howard, K. I., & Leon, S. (2002). Outcomes management, expected treatment response and severity-adjusted provider profiling in outpatient psychotherapy. *Journal of Clinical Psychology, 58*(10), 1291–1304.

Maling, M. S., Gurtman, M. B., & Howard, K. I. (1995). The response of interpersonal problems to varying doses of psychotherapy. *Psychotherapy Research, 5,* 63–75.

Matsumoto, K., Jones, E., & Brown, J. (2003). Using clinical informatics to improve outcomes: A new approach to managing behavioural healthcare services. *The Journal on Information Technology in Healthcare, 1,* 135–150.

Mellor-Clark, J., Barkham, M., Connell, J., & Evans, C. (1999). Practice-based evidence and need for a standardised evaluation system: Informing the design of the CORE system. *European Journal of Psychotherapy, Counselling and Health, 3,* 357–374.

Merrell, K. W. (2001). *Helping students overcome depression and anxiety: A practical guide.* New York: Guilford Press.

Miller, I. J. (1996). Managed care is harmful to outpatient mental health services: A call for accountability. *Professional Psychology: Research and Practice, 27,* 349–363.

Mirin, S., & Namerow, M. (1991). Why study treatment outcome? *Hospital and Community Psychiatry, 42*, 1007–1013.

Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice, 2*, 1–27.

Mordock, J. B. (2000). Outcome assessment: Suggestions for agency practice. *Child Welfare, 79*, 689–710.

Moses-Zirkes, S. (1994, March). Outcome research: Everybody wants it. *American Psychological Association Monitor.*

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Murphy, J. M., & Jellinek, M. (1990). The recognition of psychosocial disorders in pediatric office practice: The current status of the pediatric symptom checklist. *Developmental & Behavioral Pediatrics, 11*(5), 273–278.

National Advisory Mental Health Council. (2001). *Blueprint for change: Research on child and adolescent mental health. A report by the National Advisory Mental Health Council's Workgroup on Child and Adolescent Mental Health Intervention Development and Deployment.* Bethesda. MD: National Institutes of Health/National Institute of Mental Health.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

O'Donahue, W., Graczyk, P. A., & Yeater, E. A. (1998). Quality control and the practice of clinical psychology. *Applied and Preventive Psychology, 7*, 181–187.

Pagano, M. E., Cassidy, L. J., Little, M., Murphy, J. M., & Jellinek, M. S. (2000). Identifying psychosocial dysfunction in school-age children: The Pediatric Symptom Checklist as self-report measure. *Psychology in the Schools, 37*(2), 91–106.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician, 41*(4), 311–313.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician, 44*(1), 26–30.

Pekarik, G., & Stephenson, L. A. (1988). Adult and child client differences in therapy dropout research. *Journal of Clinical Child Psychology, 17*, 316–321.

Perepletchikova, F., & Kazdin, A. E. (2005). Oppositional defiant disorder and conduct disorder. In K. Cheng & K. M. Myers (Eds.), *Child and adolescent psychiatry: The essentials* (pp. 73–88). Philadelphia: Lippincott Williams & Wilkins.

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

Richardson, L. M. & Austad, C. S. (1991). Realities of mental health practice in managed care settings. *Professional Psychology: Research and Practice, 22*, 52–59.

Ringel, J. S., & Sturm, R. (2001). National estimates of mental health utilization and expenditures for children in 1998. *Journal of Behavioral Health Services & Research, 28*, 319–333.

Rossi, P. H., Schuerman, J. R., & Budde, S. (1996). *Understanding child maltreatment decisions and those who make them. Final report of the understanding placement decisions in child welfare study.* University of Chicago, IL: Chopin Hall Center for Children.

Sabin, J. E. (1991). Clinical skills for the 1990's: Six lessons from HMO practice. *Hospital and Community Psychiatry, 42*, 605–608.

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

Schepank, H. H. (1995). *Der Beeintrdchtigungs-Schwere-Score*. Gottingen, Germany: Beltz Test Verlag.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The consumer reports study. *American Psychologist, 50*, 965–974.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92*, 581–604.

Sharfstein, S. S., & Stoline, A. M. (2000). Challenges to the preservation of quality in cost-contained behavioral health systems. In G. Stricker, & W.G. Troy (Eds.), *Handbook of quality management in behavioral health: Issues in the practice of psychology* (pp. 15–29). New York, NY: Kluwer Academic/Plenum Publishers.

Shirk, S. R., & Russell, R. L. (1992). A reevaluation of estimated of child therapy effectiveness. *Journal of American Academy of Child and Adolescent Psychiatry, 31,* 703–709.

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

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

Sperry, L., Brill, P. L., Howard, K. I., & Grissom, G. R. (1996). *Treatment outcomes in psychotherapy and psychiatric interventions.* New York: Brunner/Mazel.

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.

Steenbarger, B. N., & Smith, H. B. (1996). Assessing the quality of counseling services: Developing accountable helping systems. *Journal of Counseling and Development, 75*, 145–150.

Tang, T. Z., & DeRubeis, R. J. (1999a). Reconsidering rapid early response in cognitive behavioral therapy for depression. *Clinical Psychology: Science and Practice, 6,* 283–288.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81,* 209–219.

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

Venable, W. M. & Thompson, B. (1998). Caretaker psychological factors predicting premature termination of children's counseling. *Journal of Counseling and Development, 76*(3), 286–293.

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73,* 914–923.

Warren, J. D. & Nelson, P. L., & Burlingame, G. M. (2009). Identifying youth at risk for treatment failure in outpatient community mental health services. *Journal of Child and Family Studies, 18,* 690–701.

Warren, J. D. & Nelson, P. L., Mondragon, S. A., Baldwin, S. A., & Burlingame, G. M. (2010). Youth psychotherapy change trajectories and outcomes in usual care: Community mental health vs. managed care settings. *Journal of Consulting and Clinical Psychology, 78,* 144–155.

Weisz, J. R. (2004). *Psychotherapy for children and adolescents: Evidence-based treatments and case examples*. Cambridge: Cambridge University Press.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between lab and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63*, 688–701.

Weisz, J. R., & Gray, J. S. (2008). Evidence-based psychotherapy for children and adolescents: Data from the present and a model for the future. *Child and Adolescent Mental Health, 13*, 54–65.

Weisz, J. R., Jensen, A. L., & McLeod, B. D. (2005). Development and dissemination of child and adolescent psychotherapies: Milestones, methods, and a new deployment-focused model. In E. D. Hibbs & P. S. Jensen (Eds.), *Psychosocial Treatments for child and adolescent disorders: Empirically-based approaches* (2nd ed., pp. 9–39). Washington, DC: American Psychological Association.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

Weisz, J. R., Weiss, B., Han, S. S., Grandger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescent revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin*, *117*, 450–468.

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*(1), 59–68.

Wierzbicki, M., & Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice, 24*, 190–195.

Wilson, G. T. (1999). Rapid response to cognitive behavior therapy. *Clinical Psychology: Science and Practice, 6,* 289–292.