5-2004

A Comparison of Rational Versus Empirical Methods in the Prediction of Psychotherapy Outcome

Glen I. Spielmans
Utah State University

A COMPARISON OF RATIONAL VERSUS EMPIRICAL METHODS IN THE PREDICTION OF PSYCHOTHERAPY OUTCOME

by

Glen I. Spielmans

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR IN PHILOSOPHY

ın

Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

2004

ABSTRACT

A Comparison of Rational Versus Empirical Methods in the

Prediction of Psychotherapy Outcome

by

Glen I. Spielmans, Doctor of Philosophy

Utah State University, 2004

Major Professor: Dr. Kevin S. Masters
Department: Psychology

Several systems have been designed to monitor psychotherapy outcome, in which feedback is generated based on how a client's rate of progress compares to an expected level of progress. Clients who progress at a much lesser rate than the average client are referred to as signal-alarm cases. Recent studies have shown that providing feedback to therapists based on comparing their clients' progress to a set of rational, clinically derived algorithms has enhanced outcomes for clients predicted to show poor treatment outcomes. Should another method of predicting psychotherapy outcome emerge as more accurate than the rational method, this method would likely be more useful than the rational method in enhancing psychotherapy outcomes. The present study compared the rational algorithms to those generated by an empirical prediction method generated through hierarchical linear modeling. The sample consisted of 299 clients seen at a university counseling center and a psychology training clinic. The empirical method was significantly more accurate in predicting outcome than was the rational method. Clients predicted to show poor treatment outcome by the empirical method showed, on average, very little positive change. There was no difference between the methods in the ability to accurately forecast reliable worsening during treatment. The rational method resulted in a high percentage of false alarms, that is, clients who were predicted to show poor treatment response but in fact showed a positive treatment outcome. The empirical method generated significantly fewer false alarms than did the rational method. The empirical method was generally accurate in its predictions of treatment success, whereas the rational method was somewhat less accurate in predicting positive outcomes. Suggestions for future research in psychotherapy quality management are discussed.

(109 pages)

ACKNOWLEDGMENTS

I would like to thank my committee members (Drs. Kevin S. Masters, Michael Lambert, M. Scott DeBerard, Gretchen A. Gimpel, and David Stein) for their support and guidance throughout the process of completing this dissertation.

I also thank my family and friends, especially my father, John Spielmans, and my wife, Kara Spielmans, for their dedication and support as I worked to achieve completion of this project as well as many other important life endeavors.

Glen I. Spielmans

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

PROBLEM STATEMENT

As psychotherapy progresses into the 21st century, research has accumulated indicating that it is a potent treatment for a variety of psychological disorders (Lambert & Bergin, 1994; Smith, Glass, & Miller, 1980). The "talking cure" has been used across a variety of disorders and problems. Psychotherapy is frequently utilized in the treatment of anxiety disorders and depression, which often co-occur. Treatment for these two classes of disorders, along with treatment of substance dependence and abuse, accounts for the majority of treated psychological disorders in this country (Howard et al., 1996). Research on the effects of psychotherapy has generally undertaken three forms, all of which will be briefly discussed followed by more in-depth discussion on each.

The most popular form of psychotherapy research is on the efficacy of psychotherapy. Efficacy research relies on the use of clinical trials, which, increasingly, attempt to test the utility of a specific psychotherapy for a specific disorder. Meta-analytic (Quality Assurance Project, 1983; Smith et al., 1980; Wampold et al., 1997) and narrative (Lambert & Bergin, 1994) reviews have indicated that psychotherapy is more efficacious than both no treatment placebo treatments (Grissom, 1996; Lambert & Bergin).

Given that psychotherapy has proven generally efficacious, some researchers have compared the efficacy of one method versus another in comparative trials. Through this process, better psychotherapies should emerge as superior to lesser therapies, which would allow for the betterment of psychotherapy in general. However, these attempts have done little to prove the efficacy of one treatment over another, leaving the door open to other means of improving the outcome for psychotherapy clients (Wampold et al., 1997).

In studies of effectiveness, psychotherapy clients are followed and assessed as treatment progresses to examine the effects of treatment under realistic conditions. Results may be more applicable to clinical practice, as client populations more like those actually seen in clinical practice can be utilized with therapists providing treatments as they are actually practiced (Seligman, 1995; Shadish et al., 1997).

Another form of psychotherapy research has been recently proposed. Client-focused research (Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, Hansen, & Finch, 2001; Lambert, Okiishi, Finch, & Johnson, 1998; Lutz, Martinovich, & Howard, 1999) is based on the idea that the most important variable for a clinician is not whether a treatment works for an average client in either a clinical trial or a naturalistic setting; rather, outcome assessment should be more greatly concerned with how a treatment is working for a given client at a given point in time.

Client-focused research involves prediction of treatment response. If it were possible to devise a method of determining which clients are likely to improve in therapy and which are unlikely to improve or to deteriorate, this method would help to guide treatment. If clinicians could be alerted to clients who are not likely to improve, or more critically, to deteriorate, a change in treatment could occur to potentially avert the negative treatment outcome (Finch, Lambert, & Schaalje, 2001; Whipple et al., 2003). One such method, an empirical examination of change scores across treatment sessions, is the subject of this study. Should this method prove useful, this would set the stage for the development of therapeutic interventions that could successfully alter what are predicted to be negative courses of treatment. While different psychotherapies have given little evidence of what may improve treatment in head-to-head trials (Wampold et al., 1997), client-focused research offers, through feedback, an effort to improve psychotherapy in a different manner than comparative trials.

The goal of the current study was to compare two methods of predicting psychotherapy outcome. One method was derived by experts in the field of psychotherapy, whereas the other was empirically derived using the methods of hierarchical linear modeling. The idea is to test which model more accurately predicts psychotherapy outcome. Researchers have studied the predictive ability of rational, clinically derived methods in various areas of clinical psychology, finding that, generally, rational methods do not predict well, and that empirical methods seem more reliable and predictive than do rational methods.

The present study will compare the ability of these two methods to predict outcome. Given that feedback on client progress in psychotherapy has been shown to enhance psychotherapy outcome, it seems prudent to ensure that the most accurate predictive feedback is being given to therapists in order to maximize the effectiveness of feedback on psychotherapy outcome. Thus, research is needed to determine which method gives the most accurate feedback to therapists in the hope that more accurate predictive feedback will provide a stronger base for clinicians to intervene in the cases where unsatisfactory outcome seems likely.

One previous study has examined this idea and found, in general accord with the clinical decision-making literature (Dawes, 1994; Grove & Meehl, 1996) that the empirical method was superior in predicting outcome to the rationally derived method (Lambert, Whipple, Bishop, & Vermeersch, 2002). This study seeks to replicate the previous research and help solidify the research base on which clinicians can be provided feedback on client progress.

Given that psychotherapy is often ineffective and that head-to-head trials have done little to improve the effectiveness of therapy, it seems prudent to find other avenues of improving treatment. The provision of feedback to therapists on client progress has been shown effective in enhancing outcomes. By examining whether a rational or empirical method is more accurate in predicting psychotherapy outcome, better feedback can be given to therapists, and the outcomes of therapy can potentially be improved.

CHAPTER II

REVIEW OF LITERATURE

Introduction

Psychotherapy has shown effectiveness with clinical trial and clinically representative populations. Despite these generally positive findings, the demonstration of overall effectiveness provides little guidance for the psychotherapist whose client is not responding to psychotherapy. Thus, a new client-focused research paradigm has been developed. This line of research focuses on how outcomes can be improved for clients who are struggling in their current course of psychotherapy. Client-focused research has developed algorithms with which clients, based on their course of progress in psychotherapy, can be identified as likely to show a negative treatment response. Providing feedback to clinicians based on these predictions of treatment failure has been effective in enhancing outcomes. However, the algorithms that have been used in these feedback studies were designed using a combination of psychometrics and clinical judgment. Previous research has indicated that, when making clinical decisions, clinical judgment is often outperformed by purely empirical methods. Thus, it seems likely that a purely empirical method would outperform a set of algorithms that combines components of both clinical and empirical methods. Should an empirical method prove superior, then its use in feedback research may help to enhance outcomes beyond the positive results seen in prior studies.

Efficacy of Psychotherapy

The "Gold Standard"

Efficacy is based on the paradigm of the randomized clinical trial (RCT) as the gold standard for treatment research. As efficacious psychopharmacological interventions were developed, the RCT became the method of choice. RCTs involve the random assignment of subjects to treatment or control conditions to eliminate preexisting between group differences and selection bias. A control or comparison group is used. These vary from a wait-list control on the less stringent end, to a placebo, to another active treatment in the most stringent trials. In an RCT of a pharmacological intervention, a double-blind procedure is typically utilized to assure that neither clinician nor client are aware of whether a drug or placebo is being administered. This serves to improve internal validity, the degree to which observed effects can be attributed to the intervention in question. Psychotherapy, of course, cannot be double-blinded (Seligman, 1995), as the clinician is aware of the psychotherapy being given. Psychological placebos are often used to increase blindness of the client to treatment condition. For example, one group may receive nondirective therapy while another group may receive the active treatment (e.g., Borkovec & Mathews, 1988).

Results are examined by comparing the means of groups for significant differences. A statistically significant difference favoring an active treatment over a control group is seen as evidence of treatment efficacy. Further, individual studies are often synthesized statistically through meta-analysis, in which aggregates means and effect size statistics are calculated to allow for an overall picture of efficacy to be painted across many studies.

RCTs became the method by which the United States Food and Drug Administration (FDA) approves of medical treatments. The FDA requires that multiple RCTs documenting efficacy of a treatment compared to a placebo be completed (Healy, 1997). This requirement spans back to the 1970s, as the FDA sought to approve only treatments that were based on empirical evidence. The FDA does not regulate psychotherapy, thus freeing psychotherapy research from conducting mandatory controlled trial research. However, as controlled trials became the gold standard in pharmaceutical research, psychotherapy studies also moved to adopt this method in order to improve scientific rigor and further legitimatize psychotherapy, relegating nonrandomized psychotherapy trials to a much less important role.

Findings of Efficacy

Meta-analysis has been used to analyze a broad spectrum of data on the general efficacy of psychotherapy in the treatment of various disorders and problems, finding that, on the whole, psychotherapy is an efficacious method of treatment (Lipsey & Wilson, 1993; Smith et al., 1980). Outside of showing general efficacy in improving client outcome, research in the psychotherapy clinical trials paradigm has increasingly followed the medical model of a specific treatment for a specific disorder. A large number of trials have been conducted on pure samples utilizing specific forms of treatment (behavioral, cognitive-behavioral, interpersonal, etc.). For depression, psychotherapy has been found efficacious, as indicated by several meta-analytic reviews

(Dobson, 1989; Robinson, Berman, & Neimeyer, 1990; Steinbrueck, Maxwell, & Howard, 1983). Similarly positive results have been found for psychotherapy in the treatment of anxiety disorders (Chambless & Gillis, 1993; Clum, 1989). For other disorders, including schizophrenia (Benton & Schroeder, 1990), and chronic mental illness (Asay, Lambert, Christensen, & Beutler, 1984), psychotherapy has also shown significant efficacy.

Comparative Trials

The efficacy research paradigm has attempted to contribute to the enhancement of psychotherapy through proving that given treatments are efficacious and by attempting to demonstrate that some forms of psychotherapy are superior to others. Wampold et al. (1997) noted that previous meta-analytic reviews of psychotherapy efficacy have occasionally found a difference favoring one form of therapy over another. However, these differences are generally uncommon, especially between therapies that are considered *bona fide*, meeting the following criteria: delivered by trained therapists, based on psychological principles, were offered to the psychotherapy community as viable treatments (such as through books or manuals), or containing specified components. Hence, these researchers analyzed 113 studies published in six important journals from 1970 and 1995, finding that there was no significant difference between therapies based on an omnibus test of 277 effects culled from the obtained studies. This finding of equivalence between therapies points to the occasional finding that one therapy outperforms another (e.g., Butler, Fennell, Robson, & Gelder, 1991) as an anomaly.

Another attempt at identifying more effective types of therapy has been attempted through dismantling studies, in which a “full” treatment is compared with a “reduced” therapy. For example, cognitive-behavioral therapy may be compared to a treatment such as behavioral therapy, which could be considered cognitive-behavior therapy (the complete treatment) minus the cognitive elements. Differences observed in an RCT comparing cognitive-behavioral therapy and behavioral therapy could then be attributed to the missing cognitive component. By observing which aspects of therapy seem particularly crucial to therapeutic outcome, therapies could then be designed to capitalize on the more powerful ingredients while reducing or eliminating the elements thought less important.

Some dismantling designs have found a beneficial effect for a combined treatment over one of its components (e.g. Butler et al., 1991). A meta-analysis of 27 dismantling studies, however, found that, in general, combined or “full” treatments are no more efficacious than components of the full treatment in question (Ahn & Wampold, 2001). Combined with the results from Wampold et al. (1997), it appears that the efficacy paradigm has done little to improve upon therapy practice, as both comparative trials and dismantling designs have provided little guidance as to what therapy, if any, may be more effective than another.

Shortcomings of Efficacy Paradigm

Importantly, efficacy research emphasizes the use of specific treatments for specific disorders. As managed care emphasizes accountability and insists on the delivery of cost-effective interventions, it makes sense that efficacy research would focus more on this type of specific outcome research, as it creates a medical metaphor of a specific course of treatment for a specific disease or disorder.

As internal validity is most important in clinical trials that emphasize the specific treatment of one disorder through one treatment, it is important that client samples are homogenous. Thus, potential subjects with comorbid disorders are often not accepted for enrollment in RCTs. In fact, a high percentage of people who apply to enroll in clinical trials are rejected, perhaps as many as five to ten for every participant enrolled (Thase, 1999). In clinical trials, the issue of the severity of disorder is also important, as potential participants may be rejected for lacking either sufficient severity or having a degree of severity that is judged as too great for the study. Given the numbers provided by Thase, it is indeed questionable how well the participant samples in clinical trials generalize to everyday treatment populations.

The use of pure samples and rigorous controls helps to ensure that internal validity is maximized. Given that a high percentage of potential participants are screened from participating in clinical trials, left unanswered by RCTs is the questions of what treatment may be most useful for those who fail to qualify for trial inclusion.

Efficacy research, which is analyzed based on the results of the average client in two or more treatment or control groups, has been criticized as having insufficient relevance to clinical practice (Goldfried & Wolfe, 1998; Parloff, 1984; Persons & Silberschatz, 1998), as it is difficult to know how well any given client conforms to the average participant from treatment efficacy studies, especially given the strict, perhaps unrealistic, homogeneity of RCT participants. Use of homogenous client populations for research as well as the inflexibility of some treatment protocols are seen by many clinicians as large barriers to generalizing efficacy research to "real world" treatment settings.

In summary, a vast array of literature attests to the utility of psychotherapeutic interventions for mental disorders (Lambert & Bergin, 1994; Lipsey & Wilson, 1993). However, the clinical trial model on which many of the findings are based has been labeled as artificial based on exclusion criteria as well as on methodology. In addition, the lack of superiority in head-to-head trials and dismantling designs has also been disappointing.

Thus, a fairly recent movement has examined how well psychotherapy has performed in clinically representative samples. Researchers using this method hope to expand on the external validity of psychotherapy research and hopefully offer more avenues to enhance the effects of psychotherapy through the study of how psychotherapy works in ecologically valid settings.

Effectiveness of Psychotherapy

Therapy in the "Real World"

The effectiveness research paradigm focuses on the effects of psychotherapy in real-life settings. Thus, rather than randomly assigning participants to control or treatment groups, participants who utilize psychotherapy services as actually delivered in practice are followed over time. This makes external validity much easier to grasp, as the populations studied are comprised of actual clients seen in actual treatment centers by practicing clinicians. Thus, findings are more likely to be generalizable than in efficacy research because both therapists and clients are presumably more representative of actual practice.

Clinically Representative Therapy

Research on clinically representative therapy includes client samples, therapists, and techniques typical of psychotherapy as generally practiced. Rather than merely surveying recipients of therapy, it may be of more use to perform experimental or quasi-experimental research using clinically representative therapy. Shadish and colleagues (Shadish et al., 1997; Shadish, Matt, Navarro, & Phillips, 2000) have conducted meta-analyses on data regarding the effects of psychotherapy in clinically representative conditions.

Shadish et al. (1997) asked authors of previous psychotherapy meta-analyses to provide information regarding studies that met various criteria of clinical representativeness. Specifically, they asked previous meta-analysts to provide information on studies that were conducted in nonuniversity settings, involved participants referred through usual clinical means as opposed to recruitment by the experimenter, and used experienced therapists. It was determined that these basic criteria represented a minimum for clinical representativeness. The impact of further criteria of clinical representativeness were also examined in terms of outcome. Shadish and colleagues' results indicated that therapy that was conducted in more clinically representative settings than typical efficacy studies was equivalent in treatment effect to findings reported in efficacy research, though they cautioned that only one study was fully clinically representative and the other 55 studies they examined had only partial relation to the everyday practice of psychotherapy.

Subsequently, Shadish et al. (2000) improved upon their earlier methods. Because the Shadish et al. (1997) study utilized reports from original authors of meta-analyses, several problems arose. The 13 meta-analysts who participated in the Shadish et al. study may have coded clinical representativeness variables inconsistently. The study also counted all manualized treatments as nonrepresentative, but such procedures as relaxation are often standardized in everyday treatment. The meta-analysts may have included results from methodologically questionable studies in their replies to the authors, and this may have also biased their conclusions. An important addition employed by Shadish et al. (2000) is the use of multiple regression methods to account for covariates that may be confounded with clinical representativeness. This use of regression will be elaborated upon further in discussing their findings.

Shadish et al. (2000) utilized 90 psychotherapy studies, including 41 of the 54 contained in the original (1997) analysis. The relationship between effect size and clinical representativeness was negative (*r* = -.29 or -.35 based on fixed and random effects models, respectively), indicating that therapy was less effective when given in clinically representative settings. In subsequent analyses, the authors determined that this finding was an artifact of self-selection bias, as nonrandomized studies tended to find that the more disturbed participants, those who were rated by clinicians or who rated themselves as more distressed, assigned themselves or were assigned to treatment conditions more often than those who were less distressed. Therapy in these nonrandomized studies often brought the mean distress measure scores of the treatment distressed control group, which results in an effect size of zero. Because of pretest differences, the effect of treatment was underestimated due to nonrandom assignment. Nonrandomized studies tended to be more clinically representative, which created an unfavorable impression of representative therapy due to the unimpressive results of nonrandomly assigned treatment conditions. The findings from this analysis run parallel to other findings across various fields (Colditz, Miller, & Mosteller, 1988; Heinsman & Shadish, 1996) that indicate the practice of nonrandom assignment often biases estimates of effect size.

While Shadish et al. (2000) deemed that psychotherapy is likely effective under clinically representative conditions, thus meeting an effectiveness research goal, they urge further research, as all but one of their studies were only partially representative of everyday clinical practice. They included ten criteria of clinical representativeness and found that many studies met only a few of them (these criteria can be found in Appendix A). More research on the effects of therapy under wholly clinically representative conditions is desired.

Conclusions Regarding Effectiveness Research

Effectiveness research is a more recent phenomenon than efficacy research, so it is not surprising that the evidence for effectiveness of psychotherapy is less convincing than evidence regarding efficacy. Evidence presented by Shadish and colleagues (Shadish et al., 1997, 2000) offers promising, if tentative, support for psychotherapy practiced under realistic conditions. More research on the effectiveness of psychotherapy has the potential of utilizing data from much larger samples of several thousand clients using electronic databases (Lambert, Huefner, & Nace, 1997) to test the real-world effects of psychotherapy, although utilizing a managed care database does not allow for control conditions and thus poses a major threat to internal validity.

Clinical Significance

Both efficacy and effectiveness data are somewhat problematic to interpret because of the manner in which data are analyzed and reported. Knowing that an average client showed a large treatment effect, in itself, testify to the clinical significance of the findings. For example, if a severely depressed person scores at three and a quarter standard deviations above the mean on a depression measure, and improves by one and a half standard deviations at the end of treatment, we can say that a large treatment effect was observed, but that this person is still experiencing significant symptoms of depression.

With this in mind, that clients may improve in a statistically significant manner but not in a clinically significant manner during the course of treatment, new methods for measuring change were clearly needed. Jacobson and colleagues (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991) developed methods for determining clinically significant change. There are two important points to consider to determine if clinically significant change has been made: (a) the magnitude of change must be statistically reliable, and (b) by the end of treatment, the client should more closely resemble a member of a functional population than a member of a dysfunctional population.

The reliable change index (RCI) was created to examine whether reliable change had occurred in therapy. Cut-off points are determined for different measures of psychopathology based on the measurement error of the instrument in question and where the dividing line between functional and dysfunctional populations is drawn. When change is determined to be reliable according to the RCI and the client's posttreatment score lies closer to the mean of the functional population than the dysfunctional population, clinically significant change has occurred (Jacobson & Truax, 1991).

Client-Focused Research

Goals

While both efficacy and effectiveness research address the question of the usefulness of psychotherapy, neither paradigm answers a fundamental question that is highly useful to a clinician: Is treatment working for a particular client at this point in time? Because both effectiveness and efficacy research examine the average change of a group of clients, alternative forms of more clinically relevant outcome research are needed.

Howard et al. (1996) called for client-focused research, which tracks the progress of individual psychotherapy clients with the goal of monitoring therapeutic gain. Several variants of client-focused research have recently been proposed (Barkham et al., 2001; Beutler, 2001; Kordy, Hannover, & Richard, 2001; Lambert, Hansen, & Finch, 2001; Leuger et al., 2001). The work done by Lambert and colleagues will be discussed, as their model is the only one to have offered means of enhancing psychotherapy outcome rather than merely predicting outcome.

History

As managed care became more common (Iglehart, 1996), the demand on health care providers to show that treatment is effective has grown. While volumes of psychotherapy efficacy trials have been completed and some limited data exist concerning the effectiveness of psychotherapy in more or less real-world settings (Shadish et al., 1997, 2000), much more research is needed concerning how psychotherapy progresses in highly ecologically valid settings. This quality assurance data can be useful in several ways.

Lambert et al. (1997) discuss how managed care settings provide an excellent means for data collection. To track outcomes, managed care companies can often be quite easily convinced to utilize outcome measures to track progress. Variables including a particular psychotherapist, demographics, diagnosis, and many more can be tracked to see their relationship to outcome. What makes this particularly attractive is the ecological validity, as real clients are being treated by practicing therapists in actual therapy clinics or hospitals. Perhaps of equal importance, data can be amassed that include sample sizes in the several thousands, as opposed to the fifty or one hundred that may be present in an efficacy trial. While internal validity is poor, as control groups are not used, data are quite readily applicable in an unquestionably ecologically valid manner.

In the past few years, data have been collected concerning the average course of recovery (Finch et al., 2001; Lambert, Whipple, et al., 2001). This informs both therapist and managed care provider as to the amount of progress that can be expected. Confidence intervals are presented in order to allow an understanding of where a given client's progress or deterioration falls compared to a normative sample. Therapists informed of clients doing exceptionally well and appearing to have recovered can more quickly move toward termination, whereas therapists can alter interventions for clients who show inadequate progress or deterioration. Indeed, research regarding informing therapists of client progress has recently been published, indicating that providing therapists with feedback regarding client progress tends to result in better outcomes (Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Whipple et al., 2003).

Thus, it appears that client-focused research may be useful in predicting outcome as well as the more important task of enhancing outcome. Several client-focused systems have been devised, but only two will be discussed in this review, as their methods are most directly relevant to the study at hand.

Empirically Derived Methods Utilizing the Outcome Questionnaire (OQ-45)

In much the same spirit as other client-focused researchers, Lambert and colleagues carved their own niche in research focusing on psychotherapy outcome tracking. Their research recently focused on not only generating expected treatment responses, but on changing the course of treatment that is failing at a given point in time (Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Whipple et al., 2003). This system of quality monitoring has been the only program at this point to utilize feedback in achieving better outcomes in psychotherapy. Thus, this research not only addresses whether therapy is working at a given point in time for a particular client, it provides feedback that is then delivered to therapists in order to alter therapy that seems to be taking an ineffective course.

This group of researchers utilized a single measure of distress, the Outcome Questionnaire (OQ-45; Lambert, Burlingame, et al., 1996; Lambert, Hansen, et al., 1996). The psychometric properties of this measure appear strong (Lambert, Burlingame et al.; Lambert, Hansen, et al.; Umphress, Lambert, Smart, Barlow, & Clouse, 1997) and will be discussed later in this paper. The OQ-45 is a 45-question self-report measure that measures overall level of client functioning. It has been shown to be sensitive to change in therapy clients while remaining unchanged in repeated administrations to nonclients (Vermeersch, Lambert, & Burlingame, 2000). The main strengths of the OQ-45 are its brevity (it takes only a few minutes to complete), its solid psychometric qualities (to be described later), and its inexpensiveness.

This research team has collected a fairly large volume of OQ-45s across various client samples. Any clinician or group of clinicians who wishes to use the OQ-45 is granted free use of the instrument, provided that the clinician or clinicians agree to send all completed OQ-45s to the Brigham Young University Psychotherapy Research Center for analysis of outcome.

Feedback Based on OQ-45 Scores

Using an outcome measure begs the researcher to determine what point demarcates excellent, satisfactory, and unsatisfactory treatment responses. The ultimate goal of psychotherapy outcome research utilizing the OQ-45 is to improve outcomes. The first step is to determine expected courses of treatment, which allows therapists to compare client progress to a standard. Then, therapists can receive feedback that places client progress into varying categories of progress, from failure to success. The effects of feedback can be measured to see if outcomes are enhanced. It appears key from this analysis of client-focused research that accurate prediction of outcome is fundamental to the final success of the model in bettering psychotherapy effects.

If therapists are to be given useful feedback regarding the progress of therapy, it is important to determine what method most accurately predicts clinical response. More accurate prediction allows for more accurate feedback, which then hopefully leads to more effective intervention by therapists.

The studies completed to date on outcome feedback based on OQ-45 progress across time have been based on rational methods of modeling "signal-alarms." Clinicians have determined the methods for determining what makes for treatment success or failure at various stages of the therapy process. The method of labeling clients as either likely responsive to treatment or as headed toward becoming treatment failures is the basic step in improving outcomes based on feedback. While rational, clinically derived methods have proven effective (Lambert, Whipple, et al., 2001; Lambert et al., 2002; Whipple et al., 2003), it is not clear whether the rational method is the most accurate predictor of clinical response, or if better methods may be developed.

Only one study has examined the question of which method is better for providing more accurate feedback to therapists based on OQ-45 data (Lambert et al., 2002). This study found that an empirical method, derived through hierarchical linear modeling (Bryk & Raudenbush, 1992), was more accurate in predicting treatment failures than the rational method utilized in feedback studies. The prior study will be replicated in this dissertation in an attempt to determine whether a rational or empirical method works better for predicting psychotherapy outcome.

It is assumed that the accuracy of the feedback was a helpful tool in bettering outcomes in the studies previously mentioned. However, the only study examining the predictive power of the rational method (Lambert, Whipple, Bishop et al., 2002), found results indicating that while the rational method was often accurate, the empirical method was generally more accurate in accurately identifying clients who deteriorated over the course of treatment. If giving therapists accurate feedback helps to improve outcome, then it would seem key to establish which method is, in fact, more predictive. For example, if the rational method falsely identifies a person as likely having a positive outcome when the empirical method accurately classifies a person as likely having a negative outcome, then the rational method could lead to deleterious feedback, where the therapist is led to believe that therapy is progressing adequately when a change in intervention is, in fact, indicated. The only previous investigation comparing the two methods found that these false positives were more likely to occur when using the rational versus the empirical methods. Of those predicted to have a positive outcome by the rational method, 19.4% had a negative outcome versus 0% for the empirical method. Thus, it is important that research addresses the issue of prediction for these two models so that clinicians can make treatment decisions based on the most valid prediction of each client's outcome.

Clinical Decision Making

Introduction

Clinicians make a multitude of decisions in the assessment and psychotherapy process. They must decide on which client symptoms are to be targeted and in which order, how to structure sessions, which assessment tools to administer and how to interpret their results, and how much progress is being made in therapy, among a litany of other decisions.

Therapists have an ever-increasing number of tools at their disposal for the assessment of psychopathology and progress in psychotherapy. Clinicians thus have their own clinical judgment combined with results on objective or projective assessments to use as a basis for making decisions regarding treatment. One question that has arisen is how much weight should be assigned to clinical judgment versus objective assessment results (e.g., Dawes, 1994). Given the present focus on an empirical method versus a clinically-derived method for predicting psychotherapy outcome, the literature on rational versus empirical models of prediction in clinical psychology will be briefly reviewed.

General Findings

It appears that, in general, clinicians are not as "expert" in making decisions as many would intuitively expect (Dawes, 1994). Clinicians have been compared to statistically derived prediction rules and have often not fared well in the comparison (Garb, 1989, 1998). One example is the use of the Minnesota Multiphasic Personality Inventory (MMPI) to differentiate between neurotic and psychotic clients. Thirteen psychologists, who were rated as "experts" on the use and interpretation of the MMPI, along with 16 clinical psychology graduate students, examined a total of 861 MMPI profiles and determined if the client described in the profile was neurotic or psychotic, based on an 11-point continuum with neurotic and psychotic at opposite poles. The results obtained by clinicians and graduate students were compared to those obtained by empirical methods that used formulas to label a client's profile as psychotic or neurotic. The profiles were those of actual psychiatric patients who had received diagnoses of either neurosis or psychosis.

The results indicated that formulas were significantly more accurate than were the judgments at predicting actual patient diagnosis (Meehl, 1959). Using a sample of 402 MMPI profiles, Meehl and Dahlstrom (1960) again showed that a statistical prediction model was more accurate than clinical interpretation of MMPI profiles. Goldberg (1965) devised a number of purely empirical models of MMPI prediction that were more accurate than clinical prediction. Further research in the area of personality assessment has found that empirical methods of predicting personality are more accurate than clinicians' judgments (Meehl, 1986; Sawyer, 1966).

Clinicians' lack of accuracy in assessment when compared to statistical formulas has been well-documented, but there are several caveats that bear note. First, experts are almost always given a very small amount of information (e.g., only the results from a single test), which is not at all indicative of daily practice in which psychologists conduct extensive interviews, use multiple assessment measures, and sometimes consult historical information and obtain collateral data (Garb, 1998). There is some evidence to support the idea that more information provided to clinicians allows for better assessment (Garb, 1984; Walters, White, & Greene, 1988). Based on a small amount of information such as a single assessment score, there seems little doubt that actuarial methods outperform clinicians in prediction of psychopathology, but these studies have not provided a realistic amount of information to clinicians, which limits the results that can be drawn from these studies (Garb, 1998).

It appears, however, that clinicians can improve their validity in personality assessment when using a formula derived from their own decision-making. Using the same data presented by Meehl (1959), it was shown that by deriving a linear regression equation from judge's ratings, validity of judgment was improved significantly, though still not to the point of that reached by actuarial prediction models (Goldberg, 1970). In the original study (Meehl, 1959), raters made predictions for each individual test. Each rater's ratings were used as criterion scores and the MMPI profiles that were given to the raters were used as input in predicting the criterion. The formula derived from each clinician's responses outperformed the clinician in prediction. Goldberg attributed this to the formula reducing human error and unreliability.

Relevance to Psychotherapy Outcome

Based on the questionable strength of clinical judgment in assessment, it seems uncertain how well clinicians could forecast psychotherapy outcome. Garb (1998) noted that "statistical-prediction rules have rarely been used to make treatment decisions" (p. 222). The few studies in the area of clinician agreement on treatment assignment seem to support the idea that rational methods are often fallacious. For example, psychiatrists have shown poor agreement as to when the use of electroconvulsive therapy (ECT) is appropriate (Hermann, Dorwart, Hoover, & Brody, 1995). Another study (Keller et al., 1986) examined the treatments received by depressed patients at five university medical centers. Differences in type of treatment utilization (psychotherapy, medication, or ECT) were unrelated to the severity of the depression and the type of treatment used was best predicted by the medical center itself; that is, the place where treatment took place was more predictive than severity of illness. Agreement among psychologists has also been shown as quite poor (close to zero) when examining assignment of clients to varying levels of care. Researchers discovered that, when deciding to assign children to one of five varying levels of care, agreement among clinicians was quite poor, even when clinicians believed they had quite adequate information about the case in question (Bickman, Karver, & Schut, 1997). Other studies have found similar results (e.g. Bickman, Karver, & Schut, 1995 as cited in Bickman et al., 1997).

This line of research indicates that clinicians, the experts in the field of psychotherapy and psychiatry, often reach variant opinions regarding which form of treatment should be assigned. However, as mentioned by Salzer, Nixon, Schut, Karver, and Bickman (1997), outcomes are the most important criteria for measuring the appropriateness of treatment assignment. If perfect reliability was obtained across professionals regarding assignment for a particular case but the treatment resulted in poor outcome, then the practical validity of the treatment assignment would be poor.

Clinicians have not received high marks for their ability to make decisions when compared to statistical models (Garb, 1998; Meehl, 1986). While this has been shown in some areas, there has been little attention paid to how well empirically based versus clinically based methods predict psychotherapy outcome. The present study seeks to extend the clinical decision-making literature to the area of psychotherapy outcome.

CHAPTER III

METHOD

Introduction

Given that feedback on client progress based on OQ-45 scores across time seems to produce enhanced outcomes in psychotherapy (Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Whipple et al., 2003), it is critical to refine a method for predicting outcome that allows for accurate feedback on client progress to be given to clinicians, who can then refine psychotherapy appropriately, in accordance with client change data. This study will compare rational and empirical methods for forecasting client change.

Procedures

Participants

Archival data was retrieved regarding psychotherapy clients from the Utah State University Counseling Center (UCC). This clinic provides outpatient psychotherapy to students of Utah State University. Clients are often treated by practicum students in at least their third year of graduate training. Other therapists include licensed psychologists, predoctoral psychology interns, and graduate assistants, who are in at least their fourth year of graduate school. All nonlicensed therapists receive weekly individual and group supervision regarding their current clients. OQ-45 data from all clients seen at the center in the academic years Fall 1998-Spring 2002 were utilized. Only clients who provided at least three OQ-45s during their course of treatment were utilized.

An additional sample was obtained from archival data at the Utah State University Psychology Community Clinic (PCC). Therapists in this clinic are all graduate practicum students who receive weekly individual and/or group supervision for their cases. Data for all clients who completed at least three OQ-45s were included. Data were collected on clients seen from academic years Fall 1997 to Summer 2002. Therapists in this study did not receive feedback based on the empirical or rational methods regarding their clients' progress.

In order to protect confidentiality, client data were coded so that the researcher did not have access to any identifying information, as each participant was identified only through a client number assigned to clients by the UCC or PCC.

Measures

The outcome measure is the OQ-45 (Lambert, Hansen, et al., 1996). This measure was used for various reasons. As previous research has addressed the question of whether empirical or rational methods better predict client outcomes (Lambert, Whipple, Vermeersch, et al., 2002), it seemed logical to attempt replication of previous results using the same measure.

The reliability of the OQ-45 appears acceptable, with internal consistency averaging .93 for both student ($n = 157$) and client ($n = 289$) samples (Lambert, Hansen et al., 1996). Test-retest reliability on the same student samples was also high, averaging .82 over a retest period of four weeks (Lambert, Hansen, et al.). As the OQ-45 was designed to measure change, it is important that OQ-45 scores: (a) are sensitive to changes that occur while clients are in treatment, and (b) show differential rates of change for a client population versus a normative sample. To assess these important issues of validity, a study was conducted in which a sample of 1,176 clients undergoing psychotherapy and 284 nonclient students took the OQ-45 on several occasions over time. The psychotherapy clients showed significantly different rates of change on the majority of OQ-45 items than did nonclients. Because the majority of individual items and the OQ-45 total score showed significantly different slopes of change between the two groups, it appears that the OQ-45 is likely a useful measure of change (Vermeersch et al., 2000). Additionally, the OQ-45 has been used to track outcome in large samples of clients, and the typical loglinear relationship has been observed (Finch et al., 2001) as has been discovered in other dose-response studies of psychotherapy utilizing different measures (e.g., Howard, Kopta, Krause, & Orlinksy, 1986).

The OQ-45 has shown good concurrent validity with other measures of psychopathology. Correlations of the OQ-45 with the Symptom Check List (SCL-R; Derogatis, 1983), Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); Zung Self-Rating Anxiety Scale (Zung, 1971); State-Trait Anxiety Inventory, State Scale (Spielberger, Gorsuch, & Lushene, 1970); SF-36 Medical Outcome Questionnaire (Ware, Kosinski, & Keller, 1994); and Friedman Well-Being Scale (Friedman, 1994) have all been high, in the range of .78 to .86.

The OQ-45 consists of three subscales, based on Lambert's (1983) conceptualization of psychopathology. The scale with the largest weight, containing 25 items, is symptom distress, which contains items related to common anxious and depressive symptoms. The second scale is labeled interpersonal relations, and contains items descriptive of interpersonal relations and interpersonal dysfunction. It contains 11 items. The final scale, social role performance, contains nine items, which are related to dysfunction in common social roles, such as work and/or school. Scores for each of the 45 items are placed on one of the three subscales. A 5-point scale is used, rating the item from "never" to "almost always." The three subscale scores are summed to obtain a total score from zero to 180.

The three subscales have been subjected to some analyses of their validity. It appears, from a factor analytic study, that the three individual subscales are so highly intercorrelated that they, in fact, represent a unitary dimension of distress and psychopathology as opposed to three individual constructs (Mueller, Lambert, & Burlingame, 1998). Thus, the use of the individual subscales appears exploratory at present, while the use of the total score is recommended practice for tracking outcomes (Lambert, Hansen, et al., 1996).

Reliable Change Index (RCI)

The OQ-45 has been subjected to analyses to determine what comprises clinically significant change (Jacobson & Truax, 1991). Using normative data from 1,353 nonclients and 1,476 clients entering treatment, the RCI appeared to be 14 points (Lambert, Hansen, et al., 1996). At this point, when 14 points of change have occurred, it can be said that change is greater than measurement error. According to the same normative study, the cutoff score on the OQ-45 is 64. When a client's score falls below 64, it is concluded that their functioning more closely approximates a functional population than a client group. Thus, if a client's score falls from 87 at intake to 60 during treatment, this is coded as clinically significant change.

Overall, due to its excellent psychometric properties, demonstrated concurrent validity, and ease of administration, the OQ-45 seems a particularly appropriate instrument to use in the monitoring of psychotherapy progress.

Statistical Procedures

The minimum criteria for entry into this study was three completed OQ-45s for each client. The presence of only two OQ-45s does not allow for prediction, as only an intake score and one further score are present, so outcome cannot be predicted.

Rationally Derived Method

The rational method is a clinically derived method for measuring client change. It was derived using a combination of clinical judgment and an understanding of the psychometric properties of the OQ-45 (Lambert, Whipple, Bishop, et al., 2002) and lumped the course of therapy into three sections, sessions 2 - 4, 5 - 9, and 10 and above. Individual clients are placed into one of four categories based on the severity of their initial OQ-45 score, under the assumption that people with different levels of initial distress will show different patterns of recovery during a course of treatment. The difference between intake OQ-45 score and the score at any given session is the measure of interest. Scores on the OQ-45 are broken into four categories based on severity.

Different types of feedback are given based on the change score and initial OQ-45 score. For example, an individual may score 74 at intake and at session 4 score 84. This would be flagged as a signal-alarm (red feedback), a likely treatment failure. In contrast, a client scoring 90 at intake and scoring 68 at session 9 would be predicted to continue improvement. Given that clients will often show differing predictions of outcome at various sessions during a course of treatment, the rule used in this study, in accordance with previous research, was that the most negative prediction of outcome is used. For example, if a client has one "red" and seven "greens" over an 8-session course of treatment, then the prediction for this client would be "red." Clients who are labeled as yellow or red are predicted to have negative treatment outcomes and are labeled as signal-alarms.

The various forms of feedback, as published in Lambert, Whipple, Bishop, et al. (2002) have been used in prior research on the effects of feedback on client outcomes. They are presented in Appendix B. A sample algorithm of how various types of feedback are determined by the rational method is provided in Appendix C.

Empirically Derived Method

This method was designed through the use of hierarchical linear modeling (HLM). A previous analysis of the OQ-45 scores of 11,492 individuals indicated that a lognormal curve appeared to approximate the general recovery curve, which allowed analysis to continue without violating assumptions of normality.

This same analysis had a large enough sample size to allow generation of expected recovery curves for 50 client groups based on their intake scores. No fewer than 220 clients comprised each of the 50 bands, which each represented about 2% of the total sample (Finch et al., 2001). Score differences as small as 1 point at intake may separate some groups near the mean whereas several points separate some groups as the tails of the distribution are approached.

What HLM essentially did in this study was generate a separate regression line and error estimate for each participant. These within-subject estimates then became dependent variables at the next stage of analysis (Speer & Greenbaum, 1995).

For the purpose of making categorical assignments of prediction, tolerance intervals are calculated around the expected course of recovery. A two-tailed 80% confidence interval is created around the expected OQ-45 score at each session. This provides a cut-off score that defines those who are responding at a rate indicative of excellent outcome (treatment response is positive and above the 80% interval) or a rate suggestive of negative outcome (treatment response is negative and beyond the 80% interval).

The next categorical assignment is based on the two-tailed, 68% confidence interval that is calculated around the expected OQ-45 score at a given session. Those whose scores deviate from this tolerance interval are falling at least one standard deviation above or below the expected treatment response.

If a client falls within the 68% tolerance interval at any session, the therapist receives green feedback indicating that treatment is progressing as expected. If the client's OQ-45 score is outside of the 68% interval but is still within the 80% confidence interval, then the client is deviating by at least one standard deviation but does not fall into the worrisome 10% who may be most likely to have negative outcomes. A yellow warning is given in these cases, indicating that some change in treatment may be needed. Should the client fall outside of the 80% tolerance interval (uppermost 10% of projected outcomes), then the therapist is given a red warning that more strongly warns that treatment change is advised.

Should the client fall on the side of tolerance intervals that indicate unusually positive change, then the therapist is alerted to this development as well. If the client's OQ-45 score is below the predicted 68% tolerance interval but above the bottom 10%, meaning that it falls between the 68% and 80% tolerance intervals, then the therapist receives white feedback, indicating that the client's progress is greater than is generally expected. Should the client's score fall at the bottom 10% of expected responses, below the 80% tolerance interval, then the therapist would receive blue feedback, stating that the client is showing a significantly more positive change than is typical. It is possible that the therapist should be wary of a "flight into health," but it is more likely that psychotherapy or other events have produced an impressive change given that rapid response to treatment is related to better long-term outcomes (Haas, Hill, Lambert, & Morrell, 2002). Table 1 contains a summary of how predictions are assigned by the empirical method. As with the rational method, individuals who receive red or yellow warnings are labeled as signal-alarms.

This study used the same empirical method as Finch et al. (2001). Individuals were compared to the expected course of recovery as determined by the large sample of Finch et al., meaning that a client with an intake OQ-45 score of 77 in this sample will be expected to follow the same course of recovery as in the previous study. This is because the previous research used a large enough sample that it appears using its expected course of outcomes makes a great deal more sense psychometrically than devising a new set of expected outcomes based on this rather small sample. A sample recovery curve is included in Appendix D (Finch, 2000).

Table 1

Feedback Generated by the Empirical Method

Type of feedback	Associated level of projected outcome
Red	Worst 10% of projected outcomes
Yellow	Between bottom 11% - 16% of projected outcomes
Green	Middle 68% of projected outcomes
White	Between top 11% - 16% of projected outcomes
Blue	Top 10% of projected outcomes

Comparison of Methods: Categorical Outcomes

The criteria for successful and unsuccessful outcomes follow from the methodology of clinically significant change (Jacobson & Truax, 1991). Positive outcome was defined as a client having achieved reliable change; that is, a change in OQ-45 score equal to or greater than 14 points lower at termination compared to intake OQ-45 score. Recovery was defined as a termination OQ-45 score less than or equal to 63 and having met the criteria for reliable change. Negative outcome was defined as a change in OQ-45 score of greater than 14 points higher at termination compared to intake. Deterioration was considered as a change in OQ-45 scores of greater than 14 units and a final OQ-45 score of higher than 63.

Given these criteria for outcomes, the predictions of the rational and empirical methods will be compared for accuracy. The number and rate of correct and incorrect classifications for each method was charted. Chi-square analyses compared the rates of true positives, true negatives, false positives, and false negatives between the two methods. The outcomes of the clients who were falsely predicted to fail was also examined to determine whether false alarms are related to differing outcomes across the different methods.

Comparison of Methods: Continuous Outcomes

In addition to the above analyses, which divided outcome into discrete categories, the OQ-45 was also used as a continuous variable. The rate of OQ-45 change was examined across different predictive categories generated by both methods. Thus, it was determined if clients labeled in any given category by the rational method showed differential change as opposed to clients labeled as in the same given category by the empirical method.

Additional Analyses

Differences between sites (UCC and PC) were examined by chi-square analysis on such variables as sex, age, and initial OQ-45 severity. For the sake of quality management at both UCC and PC, data were analyzed examining general trends of recovery at both sites.

Research Questions and Hypotheses

Research Questions

1. How great of a difference will be seen between the rational and empirical method in the accurate identification of treatment failures?

2. How great of a difference will emerge between the rational and empirical method in the identification of treatment nonfailures (how will the rates differ in identifying false negative outcomes)?

3. Will each progressively more positive prediction interval relate a greater average treatment effect? Will this effect be more pronounced for the predictions of the empirical or rational method?

Hypotheses

1. Based on results from a previous investigation (Lambert, Whipple, Bishop, et al., 2002) and the general literature on clinical decision-making (Dawes, 1994; Garb, 1998; Meehl, 1986), it was predicted that the empirical method would outperform the rational method in correctly identifying treatment failures.

2. It was predicted that the empirical method would outperform the rational method in identifying treatment nonfailures (i.e., the empirical method would have a lower rate of false negative outcomes).

3. It was also predicted that each progressively positive level prediction interval would be associated with a greater average treatment effect. This effect was predicted to be more pronounced for the empirical method.

CHAPTER IV

RESULTS

Descriptive Statistics

The sample consisted of 299 clients who had attended psychotherapy at either the Utah State University Counseling Center (UCC; $n = 216$) or Utah State University Psychology Community Clinic (PCC; $n = 83$). The sample was 74.2% female and 95.4% Caucasian. Clients in this group were seen an average of 12.9 sessions from intake until collection of final data point. Frequency statistics for key demographic variables are provided in Table 2. *T*-tests were performed to examine potential differences between sites in demographic characteristics. Two significant differences emerged. UCC clients were seen for a significantly greater number of sessions at final OQ data point than their counterparts at PCC, 13.71 versus 11.02; $t(297) = 2.47, p = .014$. PCC clients were significantly older than UCC clients, 26.81 versus 23.54; $t(297) = 3.99, p < .001$. These results can be seen in Table 3.

The exact number of clients seen at PCC and UCC from 1997 to 2002 is unavailable at this time. According to the UCC clinic secretary, who worked at UCC during each year that data were collected, an estimated 600 clients were seen at UCC over the data collection period. Thus, the data collection rate was 36%, meaning that 64% of cases seen at UCC were not included in this study. The only reason cases were excluded was if they did not have at least three OQ-45 data points.

From PCC, about 250 cases were seen over the period of data collection. Many of these cases were children. Given that the OQ-45 is designed for adults, it was not

Table 2

Sample Demographic Characteristics: Frequencies

Demographic variable	Site	*N*	Percentage of sample
Number of clients	UCC	216	72.2
	PCC	83	27.8
	Total	299	100
Client sex (female)	UCC	157	72.7
	PCC	65	78.3
	Total	222	74.2
Therapist sex (female)	UCC	114	52.8
	PCC	44	53.0
	Total	158	52.8
Client race	UCC		
	Caucasian	207	96.3
	Latino	4	1.9
	Native American	2	0.9
	Asian	1	0.5
	"International Student"	1	0.5
	Missing	1	
	PCC		
	Caucasian	64	92.8
	Latino	3	4.3
	Black	1	1.4
	Asian	1	1.4
	Missing	14	

administered to children, thus excluding children from the study. About 130 adult cases were seen in the PCC during the time period when data were collected for this study. The author of this study was formerly employed in a position that tracked data for the PCC. The estimate of 130 cases comes from the projection of previously collected PCC data from the years 1999-2002 (i.e., total number of adult clients seen from 1999-2002)

Table 3

Sample Demographic Characteristics: Means

Variable	Site	Mean	(*SD*)	Difference
Client age	UCC	23.59	(5.40)	UCC > PCC *t*(297) = 3.99,
	PCC	26.81	(8.31)	*p* <. 001, *ES* = .52
Session at final data point	UCC	13.71	(7.45)	UCC > PCC *t*(297) = 2.47,
	PCC	11.02	(10.52)	*p* = .014, *ES* = .32

onto the entire time frame of the study. The data collection rate was notably higher for PCC (63.8%) than for UCC (36%). This is unsurprising given that the UCC aims to give the OQ-45 at every third session, whereas PCC policy is to administer the OQ-45 at each session.

Degree of Improvement

Overall, clients at both sites tended to show notable improvement in OQ-45 scores over the course of psychotherapy. From an average intake score of 80.76, the average client improved by 16.67 points to a final score of 64.09. As can be seen in Table 4, there was no difference between sites in intake OQ-45 scores, though clients in PCC showed significantly lower final OQ-45 scores, $F\,(1, 297) = 6.61, p = .011$. An ANOVA showed no difference between sites in OQ change during treatment, $F\,(1, 297) = 2.14, p = .15$. However, when initial OQ-45 scores were used as a covariate, a difference in OQ-45 change between sites emerged. An ANCOVA controlling for

Table 4

Average Degree of Improvement: Means

Variable	Site	Mean	(*SD*)	Difference
Intake OQ-45	UCC	81.81	(22.40)	
	PCC	78.05	(25.27)	*F* (1, 297) = 1.57,
	Total	80.76	(23.25)	*p* = .212, *ES* = .16
Final OQ-45	UCC	66.26	(21.64)	
	PCC	58.43	(28.02)	PCC < UCC
	Total	64.09	(23.80)	*F* (1, 297) = 6.61,
				p = .011, *ES* = .33
OQ-45 change during treatment	UCC	15.11	(20.71)	
	PCC	20.74	(23.65)	PCC > UCC
	Total	16.67	(21.60)	F (1, 297) = 5.075,
				p = .025[a], *ES* = .26

[a] This analysis was calculated using intake OQ as a covariate.

intake OQ-45 severity found a significant difference showing more change among PCC clients than for UCC clients, *F* (1, 297) = 5.08, *p* = .025. The standardized mean effect size difference after adjusting for intake OQ-45 severity shows a small .26 *ES* favoring PCC clients. In sum, 52.8% of clients made reliable improvement, as defined by an improvement of 14 points or greater in OQ-45 score at endpoint. Only 16 clients (5.4% of the sample) suffered a reliable increase in distress, as defined by an increase of OQ-45 score of 14 points or greater during the course of treatment. Summary information of categorical outcomes is provided in Table 5. Of those 16 clients who showed reliable negative change, 13 deteriorated, showing an increase of OQ-45 score by at least 14 points as well as ending treatment with an OQ-45 score of at least 64. The breakdown of categorical outcomes by sites is listed in Table 5. Sites showed no significant difference

Table 5

Average Degree of Improvement: Categorical Outcomes

Outcome	Site	% of clients	Difference
Reliable improvement	UCC	50.9	
	PCC	57.8	
	Total	52.8	*t* (297) = 1.07, *p* = .29
Recovery	UCC	31.0	
	PCC	44.6	PCC > UCC, *t*(297) = 2.14, *p* = .03
	Total	34.8	
No reliable change	UCC	44.9	
	PCC	33.7	
	Total	41.8	*t*(297) = 1.76, *p* = .08
Reliable worsening	UCC	4.2	
	PCC	8.4	
	Total	5.4	*t*(297) = 1.47, *p* = .21
Deterioration	UCC	3.2	
	PCC	7.2	
	Total	4.3	*t*(297) = 1.52, *p* = .20

in terms of categorical outcomes with the exception of percentage of clients who met criteria for recovery (improvement of at least 14 OQ-45 points and a final OQ-45 score of 63 or less), in which a significantly greater proportion of PCC clients met recovery criteria than did UCC clients, 44.6% versus 31.0%, Levene's *F* for equal variances = 11.301, *p* =.001; *t*(139.321) = 2.21, *p* = .034.

Comparison of Methods: Hit Rates by Dichotomous Prediction

Of the 16 clients who were reliably worse posttreatment, the empirical method correctly predicted 13 (81.2%), whereas the rational method correctly predicted 11 cases for a hit rate of 69%. The difference between methods did not reach significance, χ^2 (1, $n = 16$) = 1.63, $p = .20$. Three clients who were reliably worse still scored in the nonclinical range on the OQ-45 at endpoint, leaving a total of 13 clients who met criteria for deterioration. Of these clients, both methods correctly predicted 10 (76.9%). While both methods accurately predicted similar numbers of treatment failures, differences emerged when looking at the rate of false positive and false negative outcomes. As can be seen in Table 6, the rational method had only a 60% hit rate in predicting positive outcomes, whereas the empirical method correctly predicted 81% of positive outcomes. This was due to the high rate of false alarms issued by the rational method, as its rate of false alarms that incorrectly predicted reliably negative outcome was slightly greater than twice that of the empirical method, a difference that reached statistical significance, χ^2 (1, $n =$ 299) = 48.03, $p < .0001$. Overall, the empirical method had a hit rate of 81% compared to only 60% for the rational method when using reliable worsening as the outcome criteria for a negative outcome. This difference in hit rates was statistically significant, χ^2 (1, $n = 299$) = 50.41, $p < .0001$.

As can be seen in Table 7, of clients predicted to fail by the empirical method, 19.4% worsened, 53.7% showed no reliable change, and only 26.9% improved reliably. In contrast, among clients predicted to show positive outcome by the empirical method, 38.4% showed no reliable change, 60.3% showed reliable improvement, and 1.3% reliably changed negatively. The difference in percentage of clients showing reliable improvement between positive and negative empirical predictions was significant, χ^2 (1, $n = 232$) = 131.97, $p < .0001$. Among clients falsely predicted to fail by the empirical

Table 6

Comparison of Hit Rates by Prediction Method: Reliable Worsening as Negative Outcome Criteria

		Predicted positive outcome		Predicted negative outcome			
	Classification method	*N*	(%)	*N*	(%)	Total	%
		Hits		False negatives			
Actual positive outcome	Rational	171	(60.4)	112	(39.6)	283	94.6
	Empirical	229	(80.9)	54	(19.1)	283	94.6
		False positives		Hits			
Actual negative outcome	Rational	5	(31.3)	11	(68.7)	16	5.4
	Empirical	3	(18.3)	13	(81.2)	16	5.4
Total number classified	Rational	176	(58.9)	123	(41.1)	299	100
	Empirical	232	(77.6)	67	(22.8)	299	100
Hit rates	Rational	182	(60.9)				
	Empirical	242	(80.9)				
Misses	Rational	117	(39.1)				
	Empirical	57	(18.1)				

method, 56% showed no reliable change while 44% made reliable positive change. Of clients who were predicted to succeed according to the rational method, 62% made reliable positive change whereas 3% showed reliable worsening and 35% showed no reliable change. The percentage of clients who improved reliably was significantly different between those who received positive versus negative predictions of outcome, χ^2 (1, n = 123) = 25.39, p < .0001. Using deterioration as the negative outcome criteria (Table 8), the difference in hit rates is virtually identical as when using reliable

Table 7

Categorical Outcomes by Signal-Alarm and Nonsignal-Alarm Predictions

Prediction	Reliably improved		No reliable change		Reliably worse	
Rational: signal-alarm	49	(39.8%)	63	(51.2%)	11	(8.9%)
Rational: not signal-alarm	109	(61.9%)	62	(35.2%)	5	(2.7%)
Empirical: signal-alarm	16	(26.9%)	36	(53.7%)	13	(19.4%)
Empirical: not signal alarm	140	(60.3%)	89	(38.4%)	3	(1.3%)

worsening as the negative outcome criteria, with a 79.9% hit rate for the empirical method versus a 61.2% hit rate for the rational method.

Whether using reliable worsening or deterioration as the criteria for negative outcome, the empirical method was significantly more accurate in making dichotomous outcome predictions (negative vs. nonnegative outcome).

Comparison of Methods: Continuous Outcomes by Dichotomous Prediction

Data on OQ-45 change was transformed into a standard format. When transformed into a standardized mean difference effect size (*ES*; intake OQ-45 score-endpoint OQ-45 score/(pooled standard deviation of intake and endpoint OQ-45 scores), those clients predicted to fail by the empirical method improved by a small *ES* of .17.

Table 8

Comparison of Hit Rates by Prediction Method: Deterioration as Negative Outcome Criteria

		Predicted positive outcome		Predicted negative outcome			
	Classification method	*N*	(%)	*N*	(%)	Total	%
		Hits		False negatives			
Actual positive outcome	Rational	173	(60.5)	113	(39.5)	286	95.7
	Empirical	229	(80.9)	57	(19.9)	286	95.7
		False positives		Hits			
Actual negative outcome	Rational	3	(23.1)	10	(76.9)	13	4.3
	Empirical	3	(23.1)	10	(76.9)	13	4.3
Total number classified	Rational	176	(58.9)	123	(41.1)	299	100
	Empirical	232	(77.6)	67	(22.8)	299	100
Hit rates	Rational	182	(61.2)				
	Empirical	242	(79.9)				
Misses	Rational	117	(38.8)				
	Empirical	57	(20.1)				

This indicates that little improvement occurred for those clients labeled as negative by the empirical method. Clients predicted to have negative outcome by the rational method improved by an average of 12 points on the OQ-45 (*ES* = .53), indicating that the average outcome for a client predicted to fail by the rational method was generally somewhat positive, showing a notable contrast to clients predicted to fail by the empirical method. Clients predicted to have a neutral or positive outcome (i.e., *not* to have a negative response to treatment) by both did similarly well (*ES* for positive prediction by empirical method = .90; *ES* for positive prediction by rational method = .88). Table 9 summarizes the above results.

Comparison of Methods: Categorical Outcomes by Prediction Subcategory

The rational method's red category caught 9 of 16 clients who showed reliable worsening, whereas its yellow category identified 2 clients who worsened. The empirical method's red category identified 12 of 16 clients who worsened and its yellow method detected 1 client who became reliably worse over the course of treatment. As predicted for both models, the majority of clients who reliably worsened were detected as signal alarms by both methods.

As can be seen in Table 10, for those categorized as red by the rational method, 15.3% worsened reliably or deteriorated, while 54.2% showed no reliable change, and 30.5% made reliable improvement. Clients categorized as yellow by the rational method reliably worsened in 3.1% of cases, made no reliable change in 48.4% of cases, and made reliable improvement in 48.4% of cases. Those clients labeled as green made reliable improvement 73.3% of the time, while showing no reliable change 24.8% of the time, and reliably changing for the worse only 2% of the time. For those clients placed in the most optimistic category, white, by the rational method 4% deteriorated, 46.7% improved reliably, and 49.3% made no reliable change.

Of clients labeled as red by the empirical method, 22.2% worsened reliably, the same percentage improved reliably, and 55.6% made no reliable change. Among clients

Table 9

Change in OQ-45 Scores by Rational or Empirical Prediction of Outcome

Method	Prediction	Mean *ES* change	Mean OQ-45 change
Rational	Negative	.53	12.03
Empirical	Negative	.17	3.79
Rational	Positive	.88	19.9
Empirical	Positive	.90	20.39

Table 10

Outcomes by Prediction Subcategories

		Reliably worse		Reliably improved		No reliable change				
Method	Category	*N*	%	*N*	%	*N*	%	*ES* Change	Total	%
Rational	Red	9	(15.3)	18	(30.5)	32	(54.2)	.36	59	19.7
	Yellow	2	(3.1)	31	(48.4)	31	(48.4)	.69	64	21.4
	Green	2	(2.0)	74	(73.3)	25	(24.8)	1.15	101	33.8
	White	3	(4.0)	35	(46.7)	37	(49.3)	.73	75	25.1
Empirical	Red	12	(22.2)	12	(22.2)	30	(55.6)	.07	54	18.1
	Yellow	1	(7.7)	6	(46.2)	6	(46.2)	.58	13	4.3
	Green	3	(1.6)	102	(54.0)	84	(44.4)	.75	189	63.2
	White	0	(0.0)	6	(75.0)	2	(25.0)	1.40	8	2.7
	Blue	0	(0.0)	32	(91.4)	3	(8.6)	1.56	35	11.7

labeled as yellow by the empirical method, 7.7% made reliable negative change, 46.2% improved reliably, and the same percentage showed no reliable change. Among clients labeled as green, 1.6% worsened reliably, 54.0% made reliable positive change, and 44.4% showed no reliable change. Among clients labeled as white, 75% improved while

25% made no reliable change. Finally, among clients labeled as blue by the empirical method, 91.4% improved reliably and 8.6% made no reliable change.

The subcategories of the empirical method made more accurate predictions than did those of the rational method. This was most notable for the red alarm, the most serious alert generated by these methods. Those identified as most likely to fail by the rational method actually showed a reliably positive change twice as often as those identified as red alarms by the empirical method.

Comparison of Methods: Continuous Outcomes by Prediction Subcategory

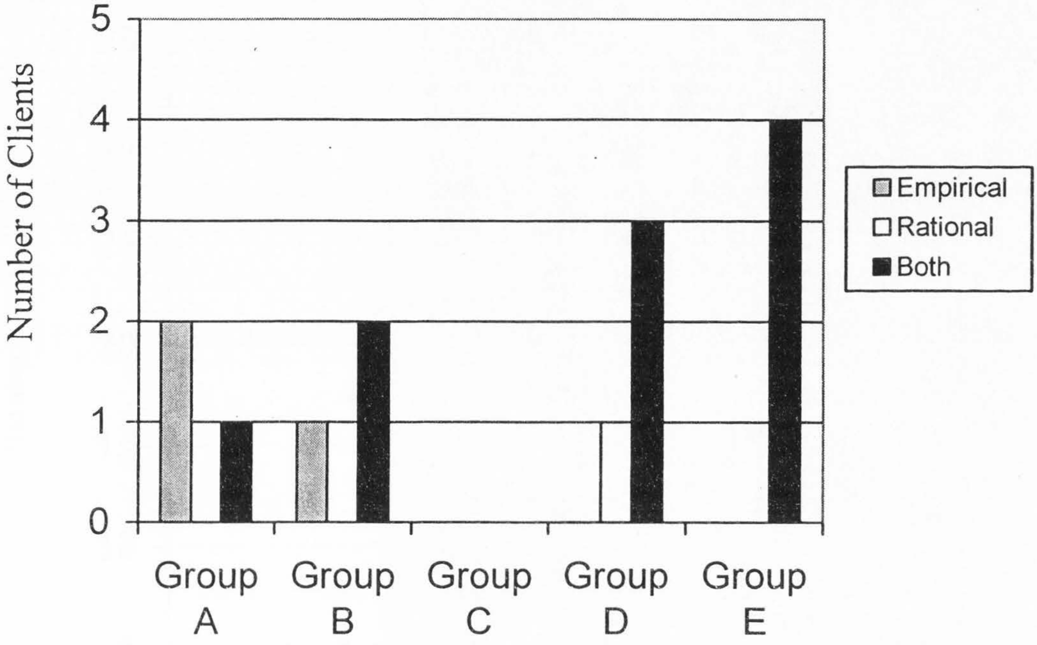
Analyses were conducted to see how much the average client changed within each subcategory of prediction for each method. The results are shown in Table 9. The average client in the red category of the rational method made small improvement (*ES* = .36), using Cohen's definition of a small *ES* (Cohen, 1988). For the rational method, those in the yellow category generally showed moderate change (*ES* = .69), and those labeled as green generally experienced notable change denoted by a large effect size (*ES* = 1.15), yet those labeled as most likely to succeed, clients in the white category showed moderate change (*ES* = .73), but less change than was observed in the green category. This result ran contrary to the hypothesis that each increasingly optimistic prediction category would yield more positive average outcomes, as the white category clients should hypothetically show the most positive results.

According to empirical predictions, clients predicted as most likely to fail in therapy, those in the red category, showed a tendency to change little during the course of treatment (*ES* = .07), whereas each increasingly optimistic prediction was related to an increased average effect of treatment: yellow (*ES* = .58), green (*ES* = .75), white (*ES* = 1.40), and blue (*ES* = 1.56).

These results show that the actual outcomes of the clients in this sample were much more in line with the predictions made by the empirical method than with those made by the rational method.

Comparison of Methods: Signal Case Detection and Signal-Alarm Generation

In accord with previous research comparing these two methods, the identification of cases who reliably worsened was broken down by degree of intake distress in order to better understand if one method outperformed another over any particular range of intake distress (Lambert, Whipple, Bishop, et al., 2002). This analysis can be seen in Figure 1. Intake severity was broken into six categories. Group A had very low severity, well below the clinical range (OQ-45 < 45). Group B had initial severity below the clinical range, whereas Group C had severity in the low clinical range (64 - 75). Group D's initial severity was in the clinical range typically seen in outpatient psychotherapy, whereas Group E (87 - 107) and especially Group F (greater than 107) reported quite high levels of initial distress. The empirical method was superior to the rational method in identifying cases at the very low (nonclinical) range of intake pathology, as well as at the very high end of initial distress (intake OQ-45 greater than 107). The rational method identified one client who worsened that was missed by the empiral method in the intake

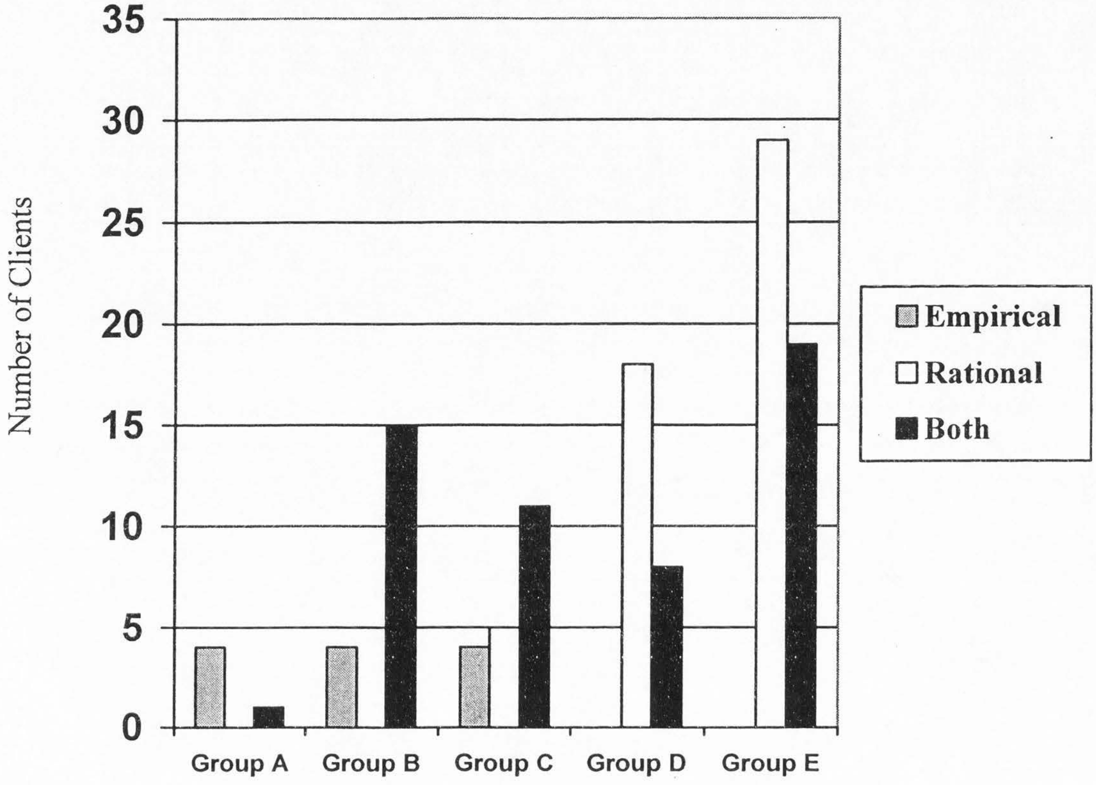


Method	A Initial scores 0-44	B Initial scores 45-63	C Initial scores 64-75	D Initial scores 76-86	E Initial scores 87-107
Rational Only	0	0	0	1	0
Empirical Only	2	1	0	0	0
Both	1	2	0	3	4

Figure 1. The relationship between degree of disturbance at intake, reliable worsening at endpoint, and identification as a single-alarm by either method or both methods jointly.

OQ-45 range of 76-86. Given the small differences etween the groups, it is difficult to interpret these findings with much certainty.

As can be seen in Figure 2, the empirical method showed a slightly greater tendency to uniquely issue a signal-alarm for clients whose intake was below the clinical range. At the higher end of intake OQ-45 scores (76 and greater), the rational method tended to predict negative outcomes at a much higher rate than did the empirical method.



Method	A Initial scores 0-44	B Initial scores 45-63	C Initial scores 64-75	D Initial scores 76-86	E Initial scores 87-107	F Initial scores 107 and above
Rational Only	0	0	5	18	29	16
Empirical Only	4	4	4	0	0	0
Both	1	15	11	8	19	1

Figure 2. The relationship between client degree of disturbance at intake and each method's pattern of signal-alarm generation.

Of the 67 signal-alarms generated by the empirical method, 54 (80.6%) were red, and 13 (19.4%) were yellow. This difference between yellow and red warnings was significant (Binomial test, $p < .0001$). For the rational method, 59 of 123 signal-alarms were red (47.9%), compared to 64 yellow alarms (52.1%); this difference was not significant.

Of the 55 cases signaled as signal-alarms by both methods, 19 (34.5%) were identified by the rational method at an earlier session than by the empirical method, while the remainder of the cases were simultaneously identified by both methods. This difference is significant based on a sign test ($z = 3.83$, $p < .0001$) and suggests that the rational method is quicker to issue alarms for cases predicted to have negative outcome. Of these cases more quickly identified by the rational method, 4 deteriorated (21.1%), 8 made reliable improvement (42.1%), and 7 (36.8%) made no reliable change.

CHAPTER V

DISCUSSION

Background

The current study was predicated on the assumption that developing clinical decision-making tools can assist practitioners in the identification of clients who are likely to not respond adequately to treatment, enabling a change in treatment plan that will hopefully lead to enhanced outcome for clients. Feedback studies have been supportive of this point, indicating that clients identified as likely to fail in treatment by an algorithm derived from a combination of the psychometric properties of the OQ-45 and expert clinical judgment have had improved outcomes when their clinicians were alerted that these clients were progressing inadequately (Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch et al., 2002; Whipple et al., 2003). An empirical, statistically derived method (Finch et al., 2001) has been developed and compared to the rational method in one previous study (Lambert, Whipple, Bishop, et al., 2002), which found the empirical method superior in detecting clients who were likely to fail during treatment. Should the empirical method emerge as consistently superior to the rational method in predicting psychotherapy outcome, then future feedback studies could implement the empirical method in the provision of feedback, enabling even greater improvement in outcome for struggling clients.

Summary of Results: Accuracy of Prediction

The present study suggests, in line with previous research (Lambert, Whipple, Bishop, et al., 2002), that the empirical method is a more accurate predictor of psychotherapy outcome than is the rational method. The empirical method identified 81% of clients who reliably worsened in treatment compared to a 69% identification rate for the rational method. Both methods accurately identified 77% of clients who deteriorated during the course of treatment. This is less than in previous research, which may be due to the OQ-45 being given somewhat infrequently to the current sample (data were collected at 49% of sessions). The previous study comparing empirical and rational methods in predicting psychotherapy outcome did not report the percentage of sessions at which OQ-45 data were collected, but it was likely much higher than in the current study. If this study did have notably fewer data points, then that would likely account for the lower identification rate of treatment failures. The rational method generated false positives at twice the rate of the empirical method, which was inconsistent with previous research that found both methods to generate false negatives at about equal levels (Lambert, Whipple, Bishop, et al.). The relatively few data points does not account for this difference, because more data points lead to more chances for a signal alarm (red or yellow warning) to be generated. Thus, the low data collection rate actually served to lower the amount of signal alarms generated.

Even with a low data collection rate, it is important to note that the empirical method was accurate in its predictions of reliable worsening versus nonworsening in 81% of cases, and in predicting deterioration versus nondeterioration in 80% of cases. This was significantly better than the rational method's accuracy rate of 60.9% and 61.2%, respectively, in identifying those who reliably worsened and who deteriorated. While both methods were nearly as accurate in identifying those who deteriorated or worsened, the rational method was responsible for generating significantly more false alarms than the empirical method, especially at higher levels of initial pathology.

Lambert, Whipple, Bishop, et al. (2002) argued that a relatively high number of false alarms is not particularly problematic when forecasting psychotherapy outcome. False positive diagnoses for many medical problems may lead to intrusive interventions and dramatic cost overruns (Northrup et al., 2002; Swets, 1992), whereas cases of psychotherapy signal-alarms merely alert the clinician to an increased likelihood of treatment failure, which can help to guide clinical interventions. The cost of false psychotherapy alarms is thus argued to be much less than the cost of false alarms for many medical diagnoses. However, therapists who are providing effective treatment may change interventions based on the receipt of false negative feedback, which could then result in the opposite of the desired effect--having therapists change from effective to ineffective interventions. Given that previous research has documented the overall effectiveness of providing feedback based on the rational method (Lambert, Whipple, et al., 2001; Lambert, Whipple, Vermeersch, et al., 2002; Whipple et al., 2003), it is likely that the benefits of altering psychotherapy due to the accurate identification of treatment failures outweigh the problem of changing effective treatment due to false negative feedback. Nonetheless, the problem of changing psychotherapy due to false negative feedback may be significant, and it is likely that the generation of less false negative feedback would lead to enhanced outcomes.

If a system consistently generates false negative feedback, as did the rational method in the current study, then its utility is limited. Therapists may grow tired of a system that quite frequently questions their clients' progress. The percentage of clients who actually show reliable negative change over the course of psychotherapy is estimated to be around 10% (Mohr, 1995). The rational method generated signal-alarms in 41% of cases, and it is likely that therapists who receive feedback indicating that such a high percentage of their clients are not responding to treatment may disbelieve or simply disregard the feedback. Given the high rate of false negatives for the rational method, the therapists' skepticism would be justified. Thus, although the provision of false negative feedback per se would quite likely not lead to untoward consequences, a system that provides too much negative feedback to therapists may simply not be accepted by therapists and thus tossed aside.

Outcome of Signal-Alarm Cases

Despite the high percentage of false negatives, it is important to note that clients who were labeled as signal-alarms showed significantly lower rates of reliable improvement during treatment. Of the 54 cases falsely predicted to become treatment failures by the empirical method, only one third showed reliable improvement, and of the 112 cases falsely predicted to fail by the rational method, only 43.8% showed reliable positive change. This indicates that even the false prediction of treatment failure is related to a decreased likelihood of positive outcome, especially when a signal-alarm is generated by the empirical method. Clients who were identified by either method as red alarms had the worst outcomes as compared to those placed into any other prediction category. The finding that 22.2% and 30.5 % of those labeled as red alarms by the empirical and rational methods, respectively, showed reliable change, is likely an artifact of the relatively low data collection rate.

Fewer data points for any individual client allow for fewer chances of a signal-alarm to be generated, so increased data collection (i.e., data collected at a higher percentage of sessions) serves to increase the number of signal-alarm cases. Thus, it is not surprising that clients labeled as likely treatment failures did somewhat better in the current study as compared to a previous investigation (Lambert, Whipple, Bishop, et al., 2002), in which only 11.6% and 14.1% of clients labeled as red made reliable positive change during treatment. However, the point made from the findings of the two studies is nonetheless clear: cases labeled as red generally show little improvement and should be taken as serious warnings that treatment is likely to result in little positive change or reliable worsening if some sort of change in intervention does not occur.

Clients labeled as yellow by either method actually made, on average, a moderate positive change during treatment. The difference in outcomes between red and yellow alarms suggests, in agreement with previous research, that a red alarm should be taken seriously as a sign that some change in treatment may be needed to improve outcome, but indicates that a yellow alarm is not nearly as troubling of a marker.

Differential Identification of Signal Cases

It appears that, similar to the previous investigation (Lambert, Whipple, Bishop, et al., 2002), the rational method was more apt to singularly identify signal alarms at the moderate to high end of the psychopathology spectrum, whereas the empirical method was more likely to uniquely identify signal cases at the lower end of the spectrum, especially in cases who presented below the clinical cut-off for clinical distress. The philosophical differences between the two methods may help explain this difference.

The rational method is especially sensitive to the identification of treatment nonresponse at the higher end of distress, under the assumption that these clients are the ones who need the most immediate reduction in their symptoms. There is some evidence supportive of the idea that more severely distressed clients who are labeled as signal cases early in treatment are more likely to conclude treatment with a negative outcome than are clients whose signal is generated later in treatment (Lambert, Whipple, et al., 2001).

The empirical method, however, makes no judgment regarding how quickly treatment should alleviate distress. This method merely provides information about how quickly a client is changing when compared to the statistically generated model of expected change. A client who presents with an OQ-45 score of 97 and scores 100 at session three would be labeled as red by the rational method, because this method assumes that a lack of progress at this point is likely to lead to deterioration because the client's distress level is rather high. The empirical method, looking at actuarial data, would generate green feedback, as it is quite typical for this level of change to have occurred between intake and the third session. It is not designed to be more sensitive to changes for clients presenting with any particular level of initial distress.

Given the high false-alarm rate of the rational method, it may not possess adequate specificity to make a strong impression on clinicians. Should a clinician be bombarded with a high percentage of warnings indicating progress is likely to be inadequate, it stands to reason that the clinician may grow tired of the high rate of negative feedback and consider it to be inaccurate. Should this occur, the utility of the alarm system would appear to be highly compromised.

Speed of Identification

When both methods labeled a case as a signal-alarm, the rational method identified about 35% of cases at an earlier session than did the empirical method, with the methods initially predicting treatment failure at the same session for the remaining 65% of cases. Of the cases that were identified earlier by the rational method as signal-alarms, 21% showed deterioration during treatment. Given that 21% is a much higher rate of deterioration than that seen in the sample as a whole, it suggests that one advantage for the rational method is its ability to predict treatment failure at an earlier session than the empirical method.

False Negative Feedback of the Rational Method

Inaccurate Algorithms

Two of the rational method's algorithms had a high propensity for the false prediction of treatment failure. For clients whose initial OQ-45 scores were higher than 72, and at session 10 or greater had shown negative change of 9 points or less, yellow feedback was generated. This subgroup of clients (n = 12) were all predicted to show reliable negative change, but none made reliably negative change, and these clients, on average, made a modest positive change at endpoint (ES = .32). This finding suggests that this particular algorithm may be overemphasizing slight negative change during the course of treatment, which then leads to the generation of false negative feedback.

For clients whose intake OQ-45 scores were 90 or above, and at session 10 or later had made between 0 and 13 points of positive change, yellow feedback was generated. Of this group (*n* = 14), none made reliable negative change, and the average client made impressively positive change at endpoint (*ES* = .77). A client who has made no change or slight positive change during the latter stages of therapy (sessions 10 and beyond) would not logically be expected to reverse course and show a reliably negative outcome. Thus, this finding suggests that the prediction of reliable worsening from clients who have made no change or are slowly making progress in treatment is inaccurate and that this particular algorithm should be revised.

If the above two changes were made, then the rational method would have made 26 fewer false negative predictions, lowering its rate of false negatives from 39.6% to 30.6%. While the latter figure is still quite high, it is certainly an improvement over the previous, unacceptably high figure.

Alterations to Feedback

The rational method tended to uniquely, and often inaccurately, label initially highly distressed cases as signal-alarms. All clients (*n* = 60) who presented with an initial OQ-45 above 72 and whose OQ-45 score is higher at any session than at intake were issued signal-alarms. Six (10%) of these clients went on to show reliable negative outcome, whereas 23 (38%) made reliable positive change. The average client in this group made moderate positive change (*ES* = .55). The generation of negative feedback by the rational method in this subgroup was useful in identifying some cases who made reliable negative change, but this negative prediction was incorrect 90% of the time.

Thus, it may be fruitful to revise the qualitative feedback given along with the yellow color code in these cases to indicate that the case appears to have a 10% chance of treatment failure, but also has a reasonably good change of succeeding in treatment, and that interventions should be monitored carefully, as opposed to providing a more negative forecast of outcome. Such revisions of the rational method may help to soften the impact of negative feedback on the clinician, especially when it is a yellow alarm. Providing actuarial data provides the clinician with a realistic assessment of the likelihood of poor outcome, which may be of greater utility than providing a blanket statement that treatment is likely to fail.

Clinical Versus Actuarial Methods

Simply stated, the results of this study support the idea that actuarial methods of prediction are generally superior to clinical methods (Garb, 1989, 1998; Grove & Meehl, 1996). The rational method, which was a hybrid of a clinical and an empirical method, was somewhat useful in predicting outcome, but was clearly outperformed by the purely empirical method.

It is important to note that this study did not directly compare clinician decision making to that of an empirical prediction model. While the empirical method can certainly be accurately labeled as an actuarial prediction model, the rational method is not a test of the judgment of individual clinicians. The rational method is a set of algorithms that uses the judgment of two experts in the field of psychotherapy, yet it is impossible to know if individual clinicians would have agreed with the various algorithms predicting likely success or failure. A more exacting test of an empirical versus a purely clinical method would have been to compare the existing empirical method with individual clinician judgment. This could be done by having clinicians, with access to the OQ-45 score from the session at hand, decide, based on this information, if treatment is progressing adequately, then code likely treatment outcome according to the various types of feedback (i.e., red, yellow). Of course, this would introduce an overwhelming confound, as therapists may indeed change their treatment based on the prediction, regardless of the prediction's actual veracity.

A more valid study could utilize blinded raters, who evaluate nothing more than the OQ-45 score at the session at hand and the intake OQ-45 score when generating judgments of likely outcome. These raters would be given normative information on the OQ-45 and could use their own clinical judgment when interpreting the difference in OQ-45 score between session OQ-45 and intake OQ-45 to determine the prediction for any given session. This study is quite likely to result in poor reliability among various raters and even within individual raters, who may well issue different predictions given equivalent amounts of change at the same session given the same intake score for different clients. Conducting such a study would likely provide a better estimate of the true difference between empirical and rational methods in predicting psychotherapy outcome.

Another study could examine the predictions of raters who watch a videotaped psychotherapy session and are also given OQ-45 scores as well as normative information for the OQ-45. This study could have respectable ecological validity, as the raters would have access to actual therapy footage as well as to OQ-45 scores, which is the same material that therapists have at their disposal. The predictions of these raters compared to those made by the empirical method would be another method for comparing the accuracy of empirical and rational methods in forecasting outcome.

In the present study, a likely more reliable form of rational prediction than that provided by a group of individual clinicians was compared with the empirical method. The methodology is somewhat similar to Goldberg (1970), who compared an empirical method to a somewhat rational method that was devised by forming a regression equation based on individual clinician guesses of psychosis versus neurosis based on MMPI profiles. Thus, three types of predictions were compared: empirical, regression based on aggregate of clinician guesses, and clinician guesses. The empirical method retained superiority, followed by the regression model, which outperformed the clinician guesses in themselves. Goldberg theorized that such a difference occurs because clinicians have fairly consistent models of prediction, but human error forces greater deviation from each person's predictive model, resulting in worse reliability for people than for purely empirical models. Put simply, while every day is the same for an empirical model, people sometimes have "off days." The reasons why individual clinicians are likely to underperform when compared to an algorithm are discussed below.

Why Clinicians Might Have Less Predictive Ability Than an Algorithm

Whether due to sleep deprivation (Pilcher & Huffcutt, 1996), various mood states (Lerner & Keltner, 2000), heuristics (Garb, 1996; Kahneman & Tversky, 1973) or confirmatory bias (Haverkamp, 1993; Pfeiffer, Whelan, & Martin, 2000), there are plenty of ways in which the clinical decision-making ability of the therapist, in this case, the ability to predict psychotherapy outcome, may be compromised on a regular basis. People are not machines; they are subject to daily variations and social psychological processes that place them at a disadvantage in comparison with a more consistent and formulaic approach to decisionmaking.

Sleep

A meta-analysis has shown that sleep deprivation negatively impacts a wide spectrum of human performance, including cognitive tasks, motor tasks, and mood (Pilcher & Huffcutt, 1996). Sleep deprivation has a very large negative impact on cognitive performance tasks, which suggests that clinicians who sleep poorly are likely to make less accurate predictions of treatment outcome. It is important to note not only chronic sleep deprivation led to decreased performance; indeed, partial sleep deprivation (less than five hours sleep in the past 24 hours) also had a large negative effect on cognitive performance. Of particular relevance to this study, research indicates that sleep-deprived medical residents perform poorer on cognitive (Eastridge et al., 2003) and surgical (Halbach, Spann, & Egan, 2003) tasks. Given that sleep problems affect an estimated 70 million Americans (National Commission on Sleep Disorders Research, 1993), lack of sleep is a likely culprit for poor performance across a number of tasks in not only research settings, but also in daily life. There is little reason to think that sleep-deprived mental health professionals would be at any lower risk for making errors under conditions of sleep deprivation than are medical residents or the population as a whole.

Mood

Evidence exists to suggest that mood state affects decision making. In general, research has indicated that positive mood state at the time of making a prediction relates to optimistic predictions whereas negative mood states are related to pessimistic predictions (Forgas, 1995). In a recent investigation that compared decisions made under two types of negative mood, people who made a decision in an angry mood were likely to make optimistic risk assessments, whereas people who make a decision in a sad mood were likely to make pessimistic risk assessments (Lerner & Keltner, 2000). This suggests that more research should be directed toward which specific emotions relate to optimistic versus pessimistic judgments. While research has not directly addressed how clinician mood impacts clinical decision making, there is no reason to believe that the prediction of psychotherapy outcome is not impacted by clinician mood at the time of prediction.

Test-Retest Reliability

Without an algorithm, the issue of reliability becomes a potential problem. Outside of the certainty that individual clinicians will interpret clinical data (including measures such as the OQ-45) differently, the question of test-retest reliability of each individual clinician's judgments arises. Each time that a clinician reviews a set of clinical data and predicts positive treatment outcome then views the same set of clinical data a week later and predicts a negative psychotherapy outcome, the predictive model's validity will suffer as a result of decreasing test-retest reliability. Research has not directly examined the reliability of clinicians' prediction of treatment outcome. The test-retest reliability of clinicians' (medical doctors and psychologists) judgment across a wider spectrum of tasks (including making diagnoses based on test data, evaluating probability and severity of disease, classifying patients into dichotomous categories) was synthesized in a meta-analytic review (Ashton, 2000). This investigation found that test-retest validity for medical doctors was .76 and .70 for psychologists. A problem with this analysis is that the interval between test and retest varied substantially between included studies, with shorter test-retest intervals generally relating to higher reliability. Results of this analysis suggest that the validity of clinical judgment is limited by temporal instability of judgment over time.

Heuristics

Heuristics refer to common guidelines that influence decisions. Since being formally identified three decades ago (Kahneman & Tversky, 1973), numerous studies have documented the existence of these decision rules that impact judgment. The representativeness heuristic refers to making a judgment based on how an object or person compares to another object or person. For example, when diagnosing depression, a therapist would be using the representativeness heuristic if he or she labeled a client as clinically depressed based on how similar a client was to what the therapist considered a "typical" case of depression. It is important to keep in mind that mental disorder diagnoses are supposedly based on whether a client meets a set of *Diagnostic and Statistical Mannual, 4th Edition* (*DSM-IV;* American Psychological Association, 1994) criteria, not on whether a client presents as "typical" of any particular diagnosis. The "typical" case of a given disorder will, of course, sometimes meet diagnostic criteria for the disorder, but will often fail to meet diagnostic criteria if the clinician does not attend to the DSM criteria. Research indicates that when making diagnoses, clinicians frequently fail to attend to DSM symptom criteria, often heavily weighing their diagnostic decisions on information that is not contained in the diagnostic criteria to the point that clinician diagnoses frequently fail to match the diagnosis as described by the symptoms in the DSM (Jampala, Sierles, & Taylor, 1988; McFall, Murburg, Smith, & Jensen, 1991; Morey & Ochoa, 1989).

In a study examining the representativeness heuristic in clinical judgment, a group of psychologists and psychology predoctoral interns examined a case history. They were asked to provide a likelihood rating that the case had one or more of four personality disorders. Participants also provided a rating describing how similar the case was to a "typical" client who has the personality disorder in question. The ratings of likelihood and typicality had a .96 correlation (Garb, 1996). In the study, 49 of 67 clinicians made an incorrect diagnosis based on the information in the case vignette. These findings suggest that clinicians may arrive at diagnostic decisions based more on their perception of typicality than of adherence to diagnostic criteria. In the prediction of psychotherapy outcome, then, clinicians may mentally weigh how similar a particular client is to a typical client who shows treatment gains and/or how similar a particular client is to a typical client who has a negative response to treatment. This use of the representativeness heuristic may lower predictive accuracy because the clinician is likely to be at least somewhat inaccurate when gauging how closely a client represents a typical treatment responder or treatment failure.

Confirmatory Bias

Confirmatory bias occurs when a person formulates an intial impression, then follows up this impression by a combination of biased information search and biased information processing that both largely exclude disconfirmatory information while placing a strong emphasis on confirmatory information. A wide variety of social psychology studies have found that the confirmatory bias occurs consistently (Nickerson, 1998; Nisbett & Ross, 1981). Research has been conducted with graduate clinical and counseling psychology trainees (Haverkamp, 1993; Pfeiffer et al., 2000), as well as licensed doctoral-level therapists (Strohmer & Shivy, 1994) in which confirmatory bias was demonstrated. Therapists tended to seek information and describe clients in a way that confirmed their initial hypothesis, even when a viable alternative hypothesis was available. Confirmatory bias should serve to decrease the accuracy of clinician predictions of psychotherapy outcome; predictive accuracy is lessened because clinicians are not placing an equal amount of emphasis on each piece of relevant information.

It is, of course, possible that some individual clinicians may be better able to predict psychotherapy outcome than does the empirical method. However, previous literature on the subject suggests that, in aggregate, it is far more likely that the empirical method would be equivalent to or more accurate than clinician predictions in predictive accuracy. A meta-analysis of the psychological and medical literature found that the accuracy of empirical predictions exceeded that of clinical predictions by a notable margin (Grove, Zald, Lebow, Snitz, & Nelson, 2000). While one could argue that an expert individual clinician may more accurately predict outcome than the empirical method, it seems much more prudent to rely on aggregate data that indicates that such superiority of any individual clinician is likely a chance finding (Grove & Meehl, 1996).

Limitations

Data Collection Rate

The limited data collection rate is certainly a limitation of this study. With a low data collection rate, the number of signal-alarm cases is quite likely reduced and the predictive accuracy of both methods is likely negatively impacted. The most negative outcome prediction was used as the final outcome prediction for each client in this study. If a client ever received yellow or red feedback during treatment, the client was labeled as a signal-alarm. Data were only collected, on average, at 49% of sessions in this study. This means that many sessions that could have generated red or yellow feedback had no data, making it highly likely that the number of clients generating signal-alarm feedback was substantially less than would have been generated under conditions of very high data collection. For example, a hypothetical client, seen for an intake and eight subsequent sessions, could have provided data at intake and sessions 2, 3, 5, and 8. Suppose the data generated green feedback at all three sessions providing predictions (2, 3, and 5). At each session during which data were not collected, a chance to generate yellow or red feedback was potentially missed. Given that the above hypothetical case was not atypical of the current data set, it is likely that the number of signal-alarm cases in the current sample was substantially less than if a much higher rate of data collection would have been achieved.

However, outside of missing some cases that reliably worsened or deteriorated, the empirical method had a good hit rate, and its predictions were neatly related to the average effect of treatment in a linear fashion, with red cases doing poorly and blue cases, on average, doing quite well during treatment. It is possible that the rational method was affected to a greater extent by the moderately low data collection rate, though there is no reason to suspect that low data collection would hamper its accuracy any more than that of the empirical method.

While the lack of data collection is a limitation, it is possible that this is, in one way, a strength. In daily clinical practice, it is likely that administration of the OQ-45 or other regular outcome measures, is at least somewhat difficult to ensure on a regular basis. Secretarial personnel are often in charge of collecting the data, and there may be other tasks of more immediate importance that are given priority over administration of outcome measures. When a rush of clients arrive at the top of an hour, it may be difficult to ensure that each client completes an OQ-45 prior to the session. Clients sometimes arrive to session late, in which case therapists often feel pressured to spend as much productive time as possible in session, not wanting to lose another 5 or 10 minutes of valuable therapy time. Thus, the results gathered in this study may be more applicable to clinical practice in general than those generated from a study in which a very high rate of OQ-45 administration occurred.

Current Sample

Because 72.2% of clients in the current sample were from a university counseling center, the sample could well be biased toward the lower end of psychopathology and age. The rate of reliable worsening (5.4%) and deterioration (4.3%) is notably less than for a general client population in which 10% are expected to be notably worse after treatment (Mohr, 1995). It is possible that the younger, relatively well-adjusted sample could have could have been more likely to respond to treatment, or, given that most of the therapy was performed by students in a training setting, it is possible that close supervision helped to decrease the incidence of negative outcomes.

The younger, less pathological sample in this study introduces a problem of restricted range. It is likely that a comparison of these two methods using a sample more representative of the wide range of psychopathology would result in increased predictive validity for both methods, as restricted range often attenuates the relationship between dependent and independent variables. The prior statement is merely speculation and should be investigated through future research examining how well these predictive models fare in a more treatment resistant population, such as a community mental health setting.

Future Directions

The present study largely supports the previous study on the topic (Lambert, Whipple, Bishop, et al., 2002), finding that the empirical method appears to predict psychotherapy outcome with more accuracy than does the rational method. In the previous investigation, the rational method generally underperformed compared to the empirical method, but in this study, the difference between methods was of a much more notable magnitude. The empirical method accurately identified all treatment failures in the previous study, but only caught about three quarters of them in the current study. A lower rate of data collection likely accounts for much of this discrepancy. Each progressively more positive level of prediction of the empirical method corresponded to a more positive outcome for the average client. These findings in sum suggest that the empirical method should be used in future feedback studies. Using this method would allow for more accurate predictive feedback to be disseminated to therapists, who could then alter treatment appropriately.

While it is clear that giving feedback to therapists helped improve outcome for clients who were progressing inadequately, it is unclear as to what kind of feedback is most helpful in actuating improved outcome. The active ingredients in feedback remain unknown. Future studies of feedback to therapists could devise various feedback conditions and compare them to see which seems to be more effective in improving outcome. In one study (Whipple et al., 2003), some therapists were provided with information regarding the client's level of perceived social support, therapeutic alliance, and readiness for change, along with a list of possible therapeutic interventions, as part of the feedback. Clients of the therapists who received these additions did better than did clients whose therapists only received the color-coded categorical feedback. While the study supported the idea that providing therapists with multidimensional feedback on various areas of client functioning as well as some ideas for specific treatment changes may be helpful, it offered little insight into what specific modality is most effective in improving outcome.

It may also be useful to develop empirical predictive models with different cutoffs than the current model. This could be useful in accurately labeling patients who are unlikely to show a positive treatment response as opposed to those predicted to show a negative response. Feedback research could then be done to see if those predicted to show little positive change show enhanced outcomes due to therapist notification of the likelihood of nonresponse and alteration of treatment.

Regardless of what direction future feedback studies follow, it seems clear that the empirical method should be the basis of providing feedback to therapists, as it has been shown more accurate in forecasting psychotherapy outcome in the current study as well as a prior investigation (Lambert, Whipple, Bishop, et al., 2002). Feedback based on the rational method was effective in enhancing outcomes in three prior studies, and it stands to reason that using the empirical method should result in even greater gains for clients in feedback studies because empirically generated feedback is of greater predictive validity than the feedback generated by the rational method.

REFERENCES

Ahn, H., & Wampold, B. E. (2001). Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. *Journal of Counseling Psychology, 48*(3), 251-257.

American Psychiatric Asociation. (1994). *Diagnostic and statistical manual of mental disorders* (4th edition). Washington, DC: Author.

Asay, T. P., Lambert, M. J., Christensen, E. R., & Beutler, L. E. (1984). *A meta-analysis of mental health treatment outcome*. Unpublished manuscript, Brigham Young University, Department of Psychology.

Ashton, R. H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making, 13*(3), 277-294.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcome benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69*(2), 184-196.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 53-63.

Benton, M. K., & Schroeder, H. E. (1990). Social skills training with schizophrenics: A meta-analytic evaluation. *Journal of Consulting and Clinical Psychology, 58,* 741-747.

Beutler, L. E. (2001). Comparisons among quality assurance systems: From outcome assessment to clinical utility. *Journal of Consulting and Clinical Psychology, 69*(2), 197-204.

Bickman, L., Karver, M. S., & Schut, L. J. A. (1995). Consensus group agreement on level of care assignments. Unpublished raw data.

Bickman, L., Karver, M. S., & Schut, L. J. A. (1997). Clinician reliability and accuracy in judging appropriate level of care. *Journal of Consulting and Clinical Psychology, 65*(3), 515-520.

Borkovec, T. D., & Mathews, A. M. (1988). Treatment of nonphobic anxiety disorders: A comparison of nondirective, cognitive, and coping desensitization therapy. *Journal of Consulting and Clinical Psychology, 56*(6), 877-884.

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

Butler, G., Fennell, M., Robson, P., & Gelder, M. (1991). Comparison of behavior therapy and cognitive behavior therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology, 59*(1), 167-175.

Chambless, D. L., & Gillis, M. M. (1993). Cognitive therapy of anxiety disorders. *Journal of Consulting and Clinical Psychology, 61*(2), 248-260.

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

Clum, G. A. (1989). Psychological interventions versus drugs in the treatment of panic. *Behavior Therapy, 20,* 429-457.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Erlbaum.

Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Information Journal, 22,* 343-352.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

Derogatis, L. R. (1983). *The SCL-90: Administration, scoring and procedures for the SCL-90.* Baltimore: Clinical Psychometric Research.

DeRubeis, R. J., & Crits-Cristoph, P. (1998). Empirically supported individual and group psychological treatments for adult mental disorders. *Journal of Consulting and Clinical Psychology, 66*(1), 37-52.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57,* 414-419.

Eastridge, B. J., Hamilton, E. C., O'Keefe, G. E., Rege, E. V., Valentine, R. J., Jones, D. J., Tesfay, S., & Thal, E. R. (2003). Effect of sleep deprivation on the performance of the simulated laparoscopic surgical skill. *American Journal of Surgery, 186*(2), 169-174.

Finch, A. E. (2000). Psychotherapy Quality Control: The statistical generation of recovery curves for integration into an early warning system (Doctoral dissertation, Brigham Young University, 2000). *Dissertation Abstracts International, 61,* 3274.

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.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model. *Psychological Bulletin, 117*(1), 39-66.

Friedman, P. H. (1994). *Friedman well-being scale.* Redwood City, CA: Mind Garden.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4,* 641-655.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*(3), 387-396.

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment. *Professional Psychology: Research & Practice*, 27(3), 272-277.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment.* Washington, DC: American Psychological Association.

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs, 79*(9, Whole No. 602).

Goldberg, L. R. (1970). Man versus model of man: A rationale plus evidence for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422-432.

Goldfried, M. R., & Wolfe, B. E. (1998). Toward a more clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology, 66*(1), 143-150.

Grissom, R. J. (1996). The magical number .7 ± .2: Meta-meta analysis of the probability of superior outcome in comparisons involving psychotherapy, placebo, and control. *Journal of Consulting and Clinical Psychology, 64*(5), 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*(2), 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Haas, E., Hill, R. D., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy maintain gains. *Journal of Clinical Psychology, 58*(9), 1157-1172.

Halbach, M. M., Spann, C. O., & Egan, G. (2003). Effect of sleep deprivation on medical resident and student cognitive function: A prospective study. *American Journal of Obstetrics & Gynecology, 188*(5), 1198-1201.

Haverkamp, B. E. (1993). Confirmatory bias in hypothesis testing for client-identified and counselor self-generated hypotheses. *Journal of Counseling Psychology, 40*(3), 303-315.

Healy, D. (1997). *The antidepressant era.* Cambridge, MA: Harvard University Press.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods, 1*(2), 154-169.

Hermann, R. C., Dorwart, R. A., Hoover, C. W., & Brody, J. (1995). Variation in ECT use in the United States. *American Journal of Psychiatry, 152,* 869-875.

Howard, K. I., Cornille, T. A., Lyons, J. S:, Vessey, J. T., Leuger, R. J., & Saunders, S. M. (1996). Patterns of mental health service utilization. *Archives of General Psychiatry, 53,* 696-703.

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., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist, 51*(10), 1059-1064.

Iglehart, J. K. (1996). Managed care and mental health. *New England Journal of Medicine, 334*(2), 131-135.

Jacobson, N S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, applications, and alternatives. *Journal of Consulting and Clinical Psychology, 67*(3), 300-307.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Jampala, C., Sierles, F. S., & Taylor, M. A. (1988). The use of DSM-III in the United States: A case of not going by the book. *Comprehensive Psychiatry, 29(1)* 39-47.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Keller, M.B., Lavori, P. W., Klerman, G. L., Andreason, N. C., Endicottt, J., Coryell, W., Fawcett, J., Rice, J. P., & Hirschfeld, R. M. A. (1986). Low levels and lack of predictors of somatotherapy and psychotherapy received by depressed patients. *Archives of General Psychiatry, 43,* 458-466.

Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart--Heidelberg Model. *Journal of Consulting and Clinical Psychology, 69*(2), 173-183.

Lambert, M. J. (1983). Introduction to the assessment of psychotherapy outcome: Historical perspective and current issues. In Lambert, M. J., Christensen, E. R., & DeJulio, S. S. (Eds.), *The assessment of psychotherapy outcome* (pp. 3-32). New York: Wiley.

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. 143-89). New York: Wiley.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy, 3*(4), 249-258.

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*(2), 159-172.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G., Huefner, J. C., & Reisinger, C. W. (1996). *Administration and Scoring Manual for the Outcome Questionnaire (OQ 45.2)*. Wilmington, DE: American Professional Credentialing Services.

Lambert, M. J., Huefner, J. C., & Nace, D. K. (1997). The promise and problems of psychotherapy research in a managed care setting. *Psychotherapy Research, 7*(4), 321-332.

Lambert, M. J., Okiishi, J. C., Finch, A. E., & Johnson, L. D. (1998). Outcome assessment: From conceptualization to implementation. *Professional Psychology: Research and Practice, 29*(1), 63-70.

Lambert, M. J., Whipple, J. L., Bishop, M. J., & Vermeersch, D. A. (2002). Comparison of empirically derived and rationally derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*(3), 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.

Lambert, M.J., Whipple, J L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9*(2), 91-103.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on emotion and choice. *Cognition and Emotion, 14*(4), 473-493.

Leuger, 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*(2), 150-158.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181-1209.

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*(4), 571-577.

McFall, M. E., Murburg, M. M., Smith, D. E., & Jensen, C. F. (1991). An analysis of criteria used by VA clinicians to diagnose combat-related PTSD. *Journal of Traumatic Stress, 4*(1), 123-136.

Meehl, P. E. (1959). A comparison of clinicians with five statistical models of identifying psychotic MMPI profiles. *Journal of Counseling Psychology, 6,* 102-109.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370-375.

Meehl, P. E., & Dahlstrom, W. G. (1960). Objective configural rules for discriminating psychotic from neurotic MMPI profiles. *Journal of Consulting Psychology, 24,* 375-387.

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

Morey, L. C., & Ochoa, E. S. (1989). An investigation of adherence to diagnostic criteria: Clinical diagnosis of the DSM-III personality disorders. *Journal of Personality Disorders, 3*, 180-192.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment, 70*(2), 248-262.

National Commission on Sleep Disorders Research. (1993). *Wake up America: A national sleep alert.* Washington, DC: Author.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nisbett, L., & Ross, M. (1981). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Northrup J. M., Miller A. C., Nardell E., Sharnprapai S., Etkind S., Driscoll J., McGarry M., Taber H. W., Elvin P., Qualls N. L., Braden C. R. (2002). Estimated costs of false laboratory diagnoses of tuberculosis in three patients. *Emerging Infectious Diseases, 8*(11), 1264-1270.

Parloff, M. B. (1984). Psychotherapy research and its incredible credibility crisis. *Clincal Psychology Review, 4*(1), 95-109.

Persons, J., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66*(1), 125-135.

Pfeiffer, A. M., Whelan, J. P., & Martin, J. M. (2000). Decision-making bias in psychotherapy: Effects of hypothesis source and accountability. *Journal of Counseling Psychology, 47*(4), 429-436.

Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep, 19*(4), 318-326.

Quality Assurance Project. (1983). A treatment outline for depressive disorders. *Australian and New Zealand Journal of Psychiatry, 17,* 129-146.

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108*(1), 30-49.

Salzer, M. S., Nixon, C. T., Schut, J. L. A., Karver, M. S., & Bickman, L. (1997). Validating quality indicators: Quality as relationship between structure, process, and outcome. *Evaluation Review, 21*(3), 292-309.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178-200.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports Study. *American Psychologist, 50*(12), 965-974.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychotherapy under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126*(4), 512-529.

Shadish, W. R., Matt, G. E., Navarro., A. M., Siegle, G., Crits-Cristoph, P., Hazelrigg, M. D., Jorm, Lyons L. C., Nietzel, M. T., Prout, H. T., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works under clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65*(3), 355-365.

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

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63*(6), 1044-1048.

Spielberger, C.D., Gorsuch, R. L., & Lushene, R. E. (1970). *The State-Trait Anxiety Self-Evaluation Questionnaire.* Palo Alto, CA: Consulting Psychologists Press.

Steinbrueck, S.M., Maxwell, S. E., & Howard, G. S. (1983). A meta-analysis of psychotherapy and drug therapy in the treatment of unipolar depression with adults. *Journal of Consulting and Clinical Psychology, 51,* 856-863.

Strohmer, D. C., & Shivy, V. A. (1994). Bias in counselor hypothesis testing: Testing the robustness of counselor confirmatory bias. *Journal of Counseling & Development, 73*(2), 191-197.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*(4), 522-532.

Thase, M. E. (1999). How should efficacy be evaluated in randomized controlled trials of treatments for depression? *Journal of Clinical Psychiatry, 60*(Suppl. 4), 23-32.

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

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74* (2), 242-261.

Walters, G. D., White, T. W., & Greene, R. L. (1988). Use of the MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology, 56*(1), 111-117.

Ware, J., Kosinski, M., & Keller, S. D. (1994). *SF-36 Physical and Mental Health Summary Scales: A user's manual.* Boston: The Health Institute, New England Medical Center

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona-fide psychotherapies: Empirically, "All must have prizes." *Psychological Bulletin, 122*(3), 203-215.

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.

Zung, W. W. (1971). A rating instrument for anxiety disorders. *Psychosomatics, 6,* 371-379.

APPENDICES

Appendix A:

Criteria of Clinical Representativeness Used in Shadish et al. (1997, 2000)

1. Problems: More clinically representative problems are mental health or behavioral problems whereas less clinically representative problems include such treatment goals as personal growth or improving underachievement.

2. Settings: More clinically representative settings include those where treatment is typically provided, such as a mental health clinic, whereas a less clinically representative setting would be a research laboratory on a university campus.

3. Referrals: Clinically representative referrals are referred through usual clinical routes, such as primary care physicians or a family member or friend, whereas less clinically representative referrals are referred through advertisements to participate in a study.

4. Therapists: More clinically representative therapists are practicing, licensed professionals, whereas less clinically representative therapists would include graduate students or researchers who are licensed but infrequently see clients.

5. Structure: More clinically representative therapies approximate therapy as actually practiced in most settings whereas less clinically representative treatments include those which use strict manualization to a degree not typically seen in everyday practice, such as manualized dynamic therapy for depression.

6. Monitoring: More clinically representative monitoring generally means that monitoring of treatment could not influence treatment. Supervision given in a manner that

may affect therapist behavior would be not clinically representative.

7. Problem heterogeneity: More clinically representative problem heterogeneity involves therapists treating various clients with varying diagnoses or problems whereas less clinically representative treatment involves treating only clients with one particular diagnosis or problem.

8. Pretherapy training: More clinically representative pretherapy training means that therapists were not given specific training involving treatment to be used in the study.

9. Therapy freedom: More clinically representative therapy freedom means that therapists were free to use a variety of techniques in all therapy they performed. Studies that required therapists to utilize a particular, narrowly constrained, treatment were considered as poorly representative in this area.

10. Number of sessions: More clinically representative number of sessions allows for a flexible number of sessions whereas less clinically representative treatment mandates a fixed number of sessions.

Appendix B:

Feedback Given to Therapists

The various forms of feedback, as published in Lambert, Whipple, et al (2002) have been used in prior research on the effects of feedback on client outcomes. Feedback is given through a chart containing a small colored sticker that corresponded with the color type of feedback (see below), and the following written messages were also typewritten on the chart:

White Feedback--"The client is functioning in the normal range. Consider termination."

Green Feedback--"The rate of change the client is making is in the adequate range. No change in the treatment plan is recommended."

Yellow Feedback--"The rate of change the client is making is less than adequate. Recommendations: consider altering the treatment plan by intensifying treatment, shifting intervention strategies, and monitoring progress especially carefully. This client may end up with no significant benefit from therapy."

Red Feedback--"The client is not making the expected level of progress. Chances are he/she may drop out of treatment prematurely or have a negative treatment outcome. Steps should be taken to carefully review this case and decide upon a new course of action such as referral for medication or intensification of treatment. The treatment plan should be reconsidered. Consideration should also be given to presenting this client at case conference. The client's readiness for change may need to be re-assessed."

Appendix C:

A Sample Algorithm from the Rational Method

(Lambert, Whipple, Bishop et al., 2002)

Intake score	Follow-up session	Follow-up score and change score	Rule	Message
T ≥ 72, ≤ 89	2-4	Delta ≥ +10	Red	The patient is not making the expected level of progress. Chances are they may drop out of treatment prematurely or have a negative treatment outcome. Steps should be taken to carefully review this case and decide upon a new course of action, such as referral for medication or intensification of treatment. The treatment plan should be reconsidered.
		Delta ≥ 0, ≤ +9	Yellow	The rate of change the patient is making is less than adequate. Recommendation: consider altering your treatment plan by intensifying treatment, shifting intervention strategies, and monitoring progress especially carefully. This patient may end up with no significant benefit from therapy.
		All else	Green	The rate of change the patient is making is in the adequate range. No change in treatment plan is recommended based on these results.
	5-8	Delta ≥ +10	Red	The patient is clearly in need of further help but the treatment is not having the expected positive impact and is not likely to have a positive result unless a way is found to strengthen the impact of treatment.
		Delta ≥ 0, ≤ +9	Yellow	The rate of change the patient is making is less than adequate. Recommendation: consider altering your treatment plan by intensifying treatment, shifting intervention strategies, and monitoring progress especially carefully. This patient may end up with no significant benefit from therapy.
		T ≥ 64, Delta < 0	Green	The rate of change the patient is making is in the adequate range. No change in treatment plan is recommended based on these results.
		T ≤ 63, Delta ≤ -9	White	The patient is functioning in the normal range. Consider termination.

(*table continues*)

Intake score	Follow-up session	Follow-up score and change score	Rule	Message
	≥ 10	Delta ≥ +10	Red	The patient is clearly in need of further help but the treatment is not having the expected positive impact and is not likely to have a positive result unless a way is found to strengthen the impact of treatment.
		Delta ≥ 0, ≤ +9	Yellow	Serious consideration should be giving to finding other treatment options and reconsidering the treatment plan. The patient is experiencing a high level of distress and although improving somewhat is clearly in need of further help but the past treatment is not having sufficient impact.
		T ≥ 64, Delta < 0	Green	The rate of change the patient is making is in the adequate range. No change in treatment plan is recommended based on these results.
		T ≤ 63, Delta ≤ -9	White	The patient is functioning in the normal range. Consider termination.

Appendix D:

Sample Expected Recovery Curve as Generated by the Empirical Method (Finch, 2000)

Intake OQ-45 Total Score 107

Session Number	Red Warning Cutoff	YellowWarning Cutoff	EXPECTED SCORE	White Warning Cutoff	Blue Warning Cutoff
1	119	116	**106**	97	94
2	117	114	**104**	93	90
3	116	113	**102**	91	88
4	115	112	**101**	89	86
5	115	112	**100**	88	85
6	115	111	**99**	87	84
7	114	111	**98**	86	83
8	114	110	**98**	85	82
9	114	110	**97**	85	81
10	113	110	**97**	84	80
11	113	110	**97**	84	80
12	113	109	**96**	83	79
13	113	109	**96**	83	79
14	113	109	**96**	82	78
15	113	109	**95**	82	78
16	112	109	**95**	81	78
17	112	108	**95**	81	77
18	112	108	**95**	81	77
19	112	108	**94**	80	77
20	112	108	**94**	80	76

CURRICULUM VITAE

GLEN I. SPIELMANS
10238 Fairmount Dr. Apt. L
Avon, IN 46123
317-962-0972 (Work)
317-209-9878 (Home)
email: gspielma@iupui.edu

PROFESSIONAL OBJECTIVE

To obtain an academic psychology position that balances teaching and research responsibilities at an institution that emphasizes excellent undergraduate education

EDUCATION

UTAH STATE UNIVERSITY, Logan, UT (Fully APA-Accredited)
Ph.D., Clinical/Counseling/School Psychology (Adult Clinical Emphasis), August 2004 (Expected)
Grade Point Average: 3.93
Relevant Coursework in: Psychopathology, Health Psychology, Empirically Supported Treatments, Intellectual Assessment, Ethics, and Epidemiology

UNIVERSITY OF UTAH, Salt Lake City, UT
M.S., Educational Psychology, May 2001
Grade Point Average: 3.95
Relevant Coursework in: Statistics and Research Methodology, Multicultural Counseling, Group Therapy, Counseling Skills, and Personality Assessment

WESTMINSTER COLLEGE, Salt Lake City, UT
B.S., Summa Cum Laude, Psychology, May 1997
Grade Point Average: 3.94
Relevant Coursework in: Psychology and Social Sciences

RESEARCH EXPERIENCE

INDIANA UNIVERSITY SCHOOL OF MEDICINE, Indianapolis, IN
Clinical Psychology Intern, September 2003 – August 2004

- Conduct clinical interviews with research participants
- Participate in design of future studies
- Receive training on several clinical rating scales

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
Project Manager, July 2002 – July 2003
"Effects of social support and home exercise equipment on exercise adoption by individuals at risk for type 2 diabetes."
Principal Investigator: Kevin Masters, Ph.D.

- Interviewed participants regarding health and psychological states
- Designed interview protocols and health psychology interventions
- Conducted weekly health psychology interventions with participants regarding exercise behavior
- Ensured that participant recruitment goals were met
- Trained and supervised graduate and undergraduate students on interview protocols and health psychology interventions
- Entered and analyzed data
- Co-author of manuscript based on results from study
- Ordered equipment and facilitated delivery of equipment to participants

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
"A comparison of rational and empirical methods in the prediction of psychotherapy outcome."
Chairperson: Kevin Masters, Ph.D.
Doctoral Dissertation, Spring 2004 (Expected Completion Date)

- Successfully defended on March 15, 2004
- Compared two methods to predict outcome during the course of psychotherapy
- Will write up manuscript in Fall 2004 for publication

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
“Effect of a 12-week exercise program on adherence and mood in treadmill and no-treadmill groups.”
Research Assistant, August 2000 to May 2001

- Collected data on psychological and physiological measures from participants
- Entered and analyzed data
- Co-presenter of paper based on results of study

VALLEY MENTAL HEALTH: Research Unit, Salt Lake City, UT
Research Specialist, February 1999 to August 1999

- Analyzed demographic and mental health service data
- Created presentations to educate employees and other agencies
- Analyzed outcome data to assess quality of service to clients

VALLEY MENTAL HEALTH: Research Unit, Salt Lake City, UT
Drug Study Coordinator, September 1998 to January 1999

- Collected and entered patient data for studies involving psychotropic medication
- Coordinated schedules of professionals to ensure timely data collection
- Evaluated research data to ensure accuracy

PUBLICATIONS

Spielmans, G. I. (2004). Efficacy of sertraline in the treatment of children and adolescents and children with major depressive disorder: Comment. *Journal of the American Medical Association, 291 (1),* 41.

DeBerard, M. S., Spielmans, G. I., & Julka, D. (in press). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal.*

Spielmans, G. I. (2002). St. John's Wort and depression: Comment. *Journal of the American Medical Association, 288 (4),* 448- 449.

DeBerard, M. S., Masters, K. S., & Spielmans, G. I. *Psychosocial correlates of health-related quality of life in university students.* Manuscript submitted for publication.

Spielmans, G. I. *Cognitive-behavioral therapy and its components in the treatment of generalized anxiety disorder: A meta-analysis.* Manuscript submitted for publication.

Spielmans, G. I. *A critical evaluation of the monoamine theory of depression.* Manuscript submitted for publication.

Michael, K. D., Furr, R. M., Masters, K. S., Collett, B. R., & Spielmans, G. I. *Predicting clinically significant change in outpatient psychotherapy: The utility of the MMPI-2 scales.* Manuscript submitted for publication.

DeBerard, M. S., LaCaille, R. A., Spielmans, G. I., Jennings, R. D., Allen, C. A., Bentley, C. G., & Goodson, J. T. *Pre-surgical biopsychosocial variables predict long term lumbar discectomy outcomes in injured workers.* Manuscript in preparation.

Masters, K. S., Spielmans, G. I., Heath, E. M., Goodson, J. T., & Knestel, A. *The effects of mail versus telephone interventions in facilitating exercise adoption.* Manuscript in preparation.

PRESENTATIONS

Spielmans, G. I., & Mickelson, K. L. (2003, May). *Meta-analysis of cognitive and behavioral therapies for generalized anxiety disorder: Implications for efficacy and outcome assessment.* Poster presented at 83rd annual meeting of the Western Psychological Association, Vancouver, BC, Canada.

Masters, K.S., & Spielmans, G. I. (2003, May). *Religious orientation, age, and anger experience: A multivariate analysis.* Poster presented at the 83rd Annual Convention of the Western Psychological Association, Vancouver, BC, Canada.

Masters, K. S., Heath, E. M., Spielmans, G. I., Clark, K. N., & Van Langeveld, E.. (2002, May). *Effect of a 12-week exercise program on adherence and mood in treadmill and no-treadmill groups.* Paper presented at the 49th annual meeting of the American College of Sports Medicine, St.Louis, MO.

TEACHING EXPERIENCE

INDIANA UNIVERSITY PURDUE UNIVERSITY--INDIANAPOLIS: Indianapolis, IN
Instructor, Abnormal Psychology, Spring 2004

- Prepare and present course material to undergraduate students
- Assess and evaluate student progress

INDIANA UNIVERSITY PURDUE UNIVERSITY–INDIANAPOLIS: Indianapolis, IN
Instructor, Introductory Psychology, Fall 2003

- Prepare and present course material to undergraduate students
- Assess and evaluate student progress

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
Instructor, Introductory Psychology, Spring 2002, Fall 2002, Summer 2003

- Prepared and presented course material to undergraduate students
- Assessed and evaluated student progress

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
Instructor, Research Methods, Fall 2002

- Prepared and presented course material to both "live" undergraduate students and via satellite connection to students throughout Utah
- Assessed and evaluated student progress

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
Teaching Assistant, Intellectual Assessment, Fall 2001

- Corrected WAIS-III and WISC-III protocols of graduate students
- Observed and graded performance of intelligence test administrations

UNIVERSITY OF UTAH: Educational Psychology Department, Salt Lake City, UT
Instructor, Career and Life Planning, Fall 1999, Spring 2000

- Prepared and presented course material to undergraduate students
- Provided group and individual interpretation of career assessment tools

CLINICAL EXPERIENCE

INDIANA UNIVERSITY SCHOOL OF MEDICINE, Indianapolis, IN
Clinical Psychology Intern, September 2003 – August 2004

- Conduct individual psychotherapy with clients in specialty outpatient anxiety and mood disorder clinic
- Provide structured group therapy for persons with social phobia
- Perform individual psychotherapy and milieu treatment with psychiatric inpatients
- Supervise graduate students in area of child psychological assessment
- Provide psychotherapy for children and adolescents in child psychiatry clinic
- Conduct psychological assessments for adults, adolescents, and children

UTAH STATE UNIVERSITY PSYCHOLOGY COMMUNITY CLINIC, Logan, UT
Practicum Therapist, August 2000 – May 2003

- Provided individual psychotherapy to adult, adolescent, and child clients with varying concerns
- Conducted psychological assessments for all ages
- Provided parent training for parents of children with behavior disorders

UTAH STATE UNIVERSITY COUNSELING CENTER, Logan, UT
Practicum Therapist, August 2002 – May 2003

- Provided individual psychotherapy to clients presenting with different concerns
- Conducted psychological assessments
- Provided career counseling

BRIGHAM CITY COMMUNITY HOSPITAL: Cardiac Rehab Unit, Brigham City, UT
Health Psychology Practicum Therapist, May 2002 to August 2002

- Conducted intake assessments
- Designed and executed diet and exercise interventions with cardiac rehabilitation patients
- Provided individual stress management interventions

UNIVERSITY OF UTAH COUNSELING CENTER, Salt Lake City, UT
Counseling Intern, August 1999 – May 2000

- Provided individual psychotherapy
- Performed individual career counseling
- Attended weekly workshops on cultural diversity, psychotherapy and training issues

Orange Street Community Correctional Center, Salt Lake City, UT
Therapist, August 1997 – August 1998

- Conducted psychoeducational and skill development groups for mentally ill offenders

- Performed individual behavior management
- Developed psychoeducational and skill development group curricula
- Conducted intake assessments

VALLEY MENTAL HEALTH: Forensic Unit, Salt Lake City, UT
Case Manager, August 1996 – August 1997

- Facilitated utilization of community resource and entitlement programs for mentally ill offenders
- Conducted individual meetings to ensure client progress toward goals
- Co-facilitated an aftercare group focused on meeting client goals

OTHER EMPLOYMENT

UTAH STATE UNIVERSITY: Psychology Department, Logan, UT
Psychology Community Clinic Assistant, August 2001 – August 2002

- Maintained database of client data for research purposes
- Provided emergency intervention for clinic clients
- Ordered assessment materials and tracked their usage

WESTMINSTER COLLEGE: Advising Department, Salt Lake City, UT
Academic Advisor, May 1996 – June 1997

- Advised students on course selection and academic planning
- Referred students to various campus services
- Coordinated schedules of tutors

HONORS AND AWARDS

- Recipient, Presidential Fellowship, Utah State University, 2000 - 2001
- Who's Who Among Students in American Colleges and Universities, 1995 - 1997
- Treasurer, Alpha Chi National Honor Society, Westminster College Chapter, 1996 - 1997
- Westminster College Volunteer Service Award, 1994 - 1995
- Member, Alpha Chi National Honor Society

ACTIVITIES

UTAH STATE UNIVERSITY: **Student Representative,** Professional--Scientific Psychology Program, 2002 – 2004

UTAH PSYCHOLOGICAL ASSOCIATION: **Student Representative,** 2002 – 2003

JUDGE MEMORIAL CATHOLIC HIGH SCHOOL, Salt Lake City, UT: **Assistant Girls Basketball Coach,** 1999 - 2000

SPECIFIC SKILLS

- Proficient in Microsoft Office applications (Word, Excel, and PowerPoint)
- Proficient in Statistical Package for the Social Sciences (SPSS)
- Skilled at administering intellectual and personality assessments to children and adults

PROFESSIONAL MEMBERSHIP

American Psychological Association, Student Affiliate
Utah Psychological Association, Student Affiliate