

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

Glen I. Spielmans,<sup>1\*</sup> Kevin S. Masters<sup>2</sup> and Michael J. Lambert<sup>3</sup>

<sup>1</sup>Department of Psychology, State University of New York at Fredonia, Fredonia, NY, USA

<sup>2</sup>Department of Psychology, Syracuse University, Syracuse, NY, USA

<sup>3</sup>Department of Psychology, Brigham Young University, Provo, UT, USA

Recent studies have shown that providing feedback to therapists based on comparing their clients' progress to a set of rational, clinically derived algorithms delineating various categories of progress has enhanced outcomes for clients predicted to show poor treatment outcomes (signal-alarms). One prior study indicated that an empirically derived method more accurately predicts outcome than the rational method. The present study replicated the comparison of empirical and rational methods, while adding an additional layer of effect size analyses to further clarify predictive accuracy. The two methods were approximately equivalent in their accurate detection of cases that had a final negative outcome. However, the rational method had significantly lower overall predictive accuracy due to its high percentage of false negative predictions. Further, progressively more optimistic predictions based on the empirical method were associated with greater improvement for the average client. This was not the case for the rational method. Copyright © 2006 John Wiley & Sons, Ltd.

It is clear that psychotherapy generally provides substantial benefits across a fairly wide spectrum of disorders and presenting problems (Lambert & Ogles, 2004; Smith, Glass, & Miller, 1980). However, it is also clear that not all persons who receive psychotherapy reap gains (Westen & Morrison, 2001); in fact, it has been estimated that about 10% of clients deteriorate (Mohr, 1995). Managed care organizations have attempted to improve outcomes through case management procedures aiming for the early identification of clients who may fail a course of treatment (Lambert, Huefner, & Reisinger, 2000).

In a similar vein, a variety of quality management programs have been developed by psychotherapy researchers (Barkham et al., 2001; Kordy, Hannöver, & Richard, 2001; Lambert, Hansen, & Finch, 2001; Leuger et al., 2001). These programs provide various methods of tracking the progress of individual clients. While clinicians commonly monitor treatment progress, the accuracy of clinical judgment in forecasting treatment outcome is often relatively poor (Breslin, Sobell, Buchan, & Cunningham, 1997). Hannan et al. (2005) found psychotherapists particularly poor at predicting patient deterioration. In this study, therapists were provided with information that about 8% of clients in their clinic tend to deteriorate during treatment. Therapists were also familiar with the outcome measure used to define outcome in their clinic, as they had utilized the measure frequently in their practice. They were then asked to

\*Correspondence to: Glen I. Spielmans, Department of Psychology, State University of New York at Fredonia, W353 Thompson Hall, Fredonia, NY 14063, USA.  
E-mail: glen.spielmans@fredonia.edu

predict the therapeutic outcome of their current cases. Only 3 of 550 clients were predicted to deteriorate, one of whom actually deteriorated, while 39 other clients actually deteriorated during treatment who were not predicted to show such negative outcome by their therapists. Despite being armed with base rate information and having familiarity with the outcome measure utilized in the study, therapists showed an inability to accurately forecast negative outcome. These findings are in line with the general literature on clinical decision making, which has quite often found clinical prediction less accurate than predictions generated by empirical methods (Dawes, 1994; Garb, 1989). Thus, it appears likely that, on the whole, therapist judgment could benefit from decision-making support tools when assessing the progress of psychotherapy.

These support tools are, however, in their infancy in the psychotherapy research literature. In one example, researchers found four unique patterns of change during treatment for bulimia, which were able to accurately forecast outcome by the fourth week of treatment (Richard & Kordy, 2000). Follow-up research indicated that these patterns of change may prove useful in developing a treatment decision tree for clinicians that could help enhance outcomes (Hannover, Richard, Hansen, Martinovich, & Kordy, 2002). Other researchers have noted that sudden, large improvements during psychotherapy are predictive of a positive final outcome (Gaynor et al., 2003; Stiles et al., 2003), especially those that occur early in treatment (Haas, Hill, Lambert, & Morrell, 2002; Tang & DeRubies, 1999). Based on the findings that sudden improvement relates to positive outcomes, it has been suggested that poor early response to treatment should result in some form of treatment modification (Wilson, 1999).

It is possible to track client progress in psychotherapy through comparing the progress of any individual client with the expected rate of progress shown by the typical client. Clients not making adequate progress are labeled as signal-alarms. If a client is exhibiting significantly less progress than the average client, then the clinician can be alerted to this problem and alter the delivery of services by changing techniques, improving the therapeutic relationship, consulting with colleagues, or finding other appropriate action to alter the course of therapy to avert a negative outcome. Four recent studies illustrated that when therapists are alerted to poorly progressing clients, these clients tend to show enhanced final outcomes compared with

clients of therapists who are not alerted to poor client progress (Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Lambert et al., 2001b, 2002b; Whipple et al., 2003).

It is important to examine the method by which the decision to provide feedback to therapists is generated. In the aforementioned research on providing feedback to therapists, algorithms based on expert clinical judgment were utilized to generate feedback. However, when this algorithm was developed, its ability to actually predict treatment failure had not been assessed. It has since been compared with an empirically derived prediction model of treatment response (Finch, Lambert, & Schaalje, 2001), a comparison in which the empirical model more accurately, albeit more slowly, forecasted treatment failure (Lambert, Whipple, Bishop, & Vermeersch, 2002). This suggests that the empirical model may have greater accuracy in predicting negative outcomes. A better predictive model would provide the basis for the generation of more accurate feedback, allowing clients who may fail in treatment to be more accurately detected, thereby potentially averting negative outcomes.

The current study replicated the comparison between rational and empirical methods. In addition to examining categorical outcomes, the present study also investigated how the methods performed when measuring outcome on a continuous scale. It was hypothesized that, in line with prior research, both methods would generally predict treatment failure accurately, with the empirical method predicting more accurately than the rational method. It was further hypothesized that, for both methods, each progressively more positive prediction interval would be associated with increasing average gains for clients, with this trend being more pronounced for the empirical method. It was also hypothesized that the rational method would more quickly identify signal-alarm cases. Finally, we examined whether there was a tendency of either method to over- or under-predict treatment outcome based on initial level of distress.

## METHODS

### *Participants*

This study utilized participants from two sites. The first site ( $n = 216$ ) was a university counseling center. Clients at this center were often treated by practicum students ( $n = 15$ ) in at least their third

year of doctoral graduate training. Other therapists included licensed psychologists ( $n = 6$ ), predoctoral psychology interns ( $n = 10$ ), and graduate assistants ( $n = 6$ ), who were in at least their fourth year of doctoral training. All nonlicensed therapists received weekly individual and group supervision. All clients who provided at least three data points on the dependent measure during their course of treatment were utilized. These data points were required to determine the accuracy of outcome forecasts. All clients seen at the center in the consecutive academic terms Fall 1998 to Spring 2002 were eligible for participation.

An additional sample was obtained from a graduate training clinic ( $n = 83$ ), where psychotherapy was provided to community members. All 26 therapists were doctoral graduate students of varying levels of experience, who received weekly group and individual supervision from licensed psychologists regarding their current cases. The diagnoses of training clinic clients were not assessed for reliability and the use of structured interviews was rare. However, the training clinic clientele seemed reflective of the general outpatient population in that mood disorders, anxiety disorders, and relational problems were the most common concerns (personal communication with clinic director, 2004), with depression being the most common presenting problem. Further, assessment and therapy sessions were videotaped for review with supervisors; consequently, final diagnoses were determined as a result of therapist-supervisor conference. Student therapists provided a wide spectrum of psychotherapies, given the diverse theoretical orientations of the supervising faculty, though cognitive-behavioral interventions were the most commonly used. Data were collected on clients seen from academic terms 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 by the counseling center or training clinic. Clients were assigned to therapists through routine clinic procedures.

### Measures

The outcome measure was the Outcome Questionnaire (OQ-45; Lambert et al., 2004). The prior study

analyzing differential accuracy of a rational versus an empirical method in predicting psychotherapy outcome (Lambert et al., 2002a) also utilized this instrument as a dependent measure.

The reliability of the OQ-45 appears acceptable, with internal consistency in a prior study averaging 0.93 for both student ( $n = 157$ ) and client ( $n = 289$ ) samples. Test-retest reliability on the same student samples was also high, averaging 0.82 over a retest period of four weeks (Lambert et al., 2004). 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 validity issues, a study (Vermeersch, Lambert, & Burlingame, 2000) was conducted in which a sample of 1176 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 log-linear 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, & Orlinsky, 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 moderate to high, in the range of 0.53–0.86 (Lambert et al., 2004). A five-point scale is used, rating each item from 'never' (0) to 'almost always' (4). Scores for each of the 45 items are summed to obtain the total score, which ranges from zero to 180.

### 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 1353 nonclients and 1476 clients entering treatment, the RCI was determined to be 14 points (Lambert, Morton et al., 1996). Consequently, 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 clinically meaningful cut-off score on the OQ-45 is 64. When a client's score falls below 64, it is concluded that his or her functioning more closely approximates the 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 since it (a) exceeded the RCI and (b) indicates movement into the functional range. Validity data supporting the cut-off scores for clinically significant change on the OQ-45 have been reported by Beckstead, Hatch, Lambert, Eggett, Goates, and Vermeersch (2003) and Bauer, Lambert, and Nielsen (2004), as well as Lunnen and Ogles (1998).

Overall, due to its strong psychometric properties, demonstrated concurrent validity, ease of administration and use in previous studies, the OQ-45 was particularly appropriate for use in the current investigation.

### *Rationally Derived Method*

The rational method is a clinically derived method for measuring client change. It was developed using a combination of clinical judgment regarding progress during psychotherapy (e.g. that lack of progress in highly disturbed individuals is of greatest concern) and an understanding of the psychometric properties of the OQ-45 (for further details see Lambert et al., 2002a). The course of therapy was divided into three sections, sessions two through four, five through nine, and 10 and above, under the rationale that inadequate progress early in treatment is not as alarming as poor progress at later sessions. The two experts created a total of four types of classification based on client progress (i.e. current OQ-45 score relative to intake OQ-45 score).

Individual clients are placed into one of four categories based on the severity of their initial OQ-45 score, under the assumption that clients with different levels of initial distress will show different patterns of recovery during a course of treatment. The difference between the intake OQ-45 score and the score at any given session is the measure of interest in developing the prediction.

A pair of hypothetical case examples may help to illustrate the logic underlying these algorithms. An outpatient client presents with an initial OQ-45 score of 102, which is about a half standard deviation above the mean for inpatients and over a full standard deviation above the mean for outpatients. This initial score strongly suggests that the client is experiencing severe distress. Thus, any increase in the client's OQ-45 score, relative to intake score, at any point would be considered worthy of a red warning, and a positive change of less than seven points after two to four sessions would generate a yellow warning, as the raters would consider this rate of progress inadequate. A client presenting with an initial score of 83, which approximates the mean for outpatients, and shows negative change of 10 or more points (i.e., OQ-45 scores increase) relative to initial OQ-45 score at any point in treatment, would generate red feedback because this amount of patient worsening was considered clinically significant. However, if this client made positive change, but still scored in the clinical range (above 63), then the client would generate green feedback, as the raters agreed that progress for a case such as this appeared adequate.

Once the client completes the initial OQ-45 and a session of treatment, the decision rules of the rational method generate feedback, provided via a progress chart and colored dot. In the aforementioned studies regarding feedback given to therapists based on the rational method (Hawkins et al., 2004; Lambert et al., 2001b, 2002b; Whipple et al., 2003), colored dots were used as an attention-catching device. A graph, with a dot, can then be subsequently given to therapists after each session.

The rational method contains four types of feedback: red, yellow, green, and white.

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. The chances are that 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 (Lambert et al., 2002a, p. 153).

### *Empirically Derived Method*

This method was designed through the use of hierarchical linear modeling (HLM). A previous analysis of the OQ-45 scores of 11492 individuals indicated that a log-normal 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 one point at intake may separate some groups near the mean, whereas several points separate 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 et al., 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

Table 1. Feedback generated by the empirical method

Type of feedback	Associated level of projected outcome
Red	Worst 10% of projected outcomes
Yellow	In bottom 11–16%; worse projected outcomes
Green	Middle 68% of projected outcomes
White	In top 11–16%; better projected outcomes
Blue	Best 10% of projected outcomes

signal-alarms. Clients in the current study were compared with the expected course of recovery as determined by the large sample of Finch et al. (2001), meaning that a client with an intake OQ-45 score of 77 in this sample was expected to follow the same course of recovery as in the previous study.

### Data Collection and Analysis

According to protocol, clients at the training clinic (TC) were given the OQ-45 at every therapy session, whereas the protocol for OQ-45 administration at the counseling center CC was to administer the OQ-45 at every third session as well as prior to any scheduled final session. However, as will be discussed later, the rate of data collection was actually lower than suggested by the protocol of the two clinics. The dependent variables representing negative outcome were reliable worsening and deterioration. Reliable worsening was defined as a 14-point or greater negative change in OQ-45 score at the final data collection point relative to the intake session. Deterioration was defined as the presence of reliable worsening along with a final score in the clinical range (i.e. 64 or higher). Positive outcomes were labeled as reliable improvement, which entailed a 14-point or greater improvement from intake to final data point.

Given that clients often generated differing predictions of outcome at various sessions during a course of treatment, the rule used in this study, in accordance with previous research, was the use of the most negative outcome prediction. For example, if a client has one 'red' and seven 'greens' over an eight-session course of treatment, then the prediction for this client would be 'red.' Clients who were labeled as yellow or red at any point in treatment were predicted to have negative treatment outcomes and were labeled as signal-alarms. Neither therapists nor clients were provided with colored feedback nor were they given research derived progress charts regarding client progress, as the purpose of this study was to compare the differential predictive validity of the rational and empirical methods in forecasting psychotherapy outcome, not to examine the effects of feedback.

Data analysis generally consisted of chi-square and other nonparametric tests to examine the differences between methods in predictive accuracy. Additionally, effect sizes were calculated to show the amount of mean change corresponding with the varying predictions of each method.

## RESULTS

### Accuracy of Predictive Models: Categorical Outcomes

In order to examine the comparative accuracy of each method, clients' final outcome was examined in relation to the predictions made by the two methods. The results are shown in Table 2. When reliable worsening was used as the negative outcome criterion, the empirical method was accurate in 80.9% of cases compared with 60.9% of cases for the rational method. Of the 16 clients who reliably worsened, the rational method detected 11, while the empirical method identified 13. The difference in the methods' abilities to detect negative outcome was not statistically significant,  $\chi^2(1, n = 16) = 0.67, p = 0.41$ . When deterioration was used as the negative outcome criterion, both methods predicted 10 of 13 cases. The results are somewhat different, however, when examining non-negative outcomes. The rational method correctly identified 60% of cases who had a non-negative outcome (i.e. did not reliably worsen), whereas the empirical method accurately labeled 81% of non-negative outcomes. The difference in predictive accuracy was significant,  $\chi^2(1, n = 283) = 28.68, p < 0.0001$ . The superiority of the empirical method was due to the tendency of the rational method to make false negative predictions at a much greater rate than the empirical method (112 cases versus 54 cases).

The trichotomous outcome (reliable improvement, no reliable change, reliable worsening) of all cases arranged by dichotomous prediction (signal-alarm or not) for both methods is presented in Table 3. Of clients predicted to fail by the empirical method, 19.4% worsened, 53.7% showed no reliable change and only 26.9% improved reliably. Of clients falsely predicted to fail by the empirical method, one-third showed positive change while two-thirds made no reliable change. 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 ( $\chi^2(1, n = 299) = 23.38, p < 0.0001$ ), suggesting that the dichotomous empirical predictions were generally useful in predicting therapeutic success.

Among clients predicted to fail by the rational method, 39.8% improved, 51.2% showed no change and 8.9% reliably worsened. Of clients who

Table 2. Comparison of hit rates by prediction method: reliable worsening as negative outcome criteria

	Classification method	Predicted positive outcome		Predicted negative outcome		Total	%
		N	(%)	N	(%)		
Actual positive outcome	Rational	171	Hits (60.4)	112	False negatives (39.6)	283	94.6
	Empirical	229	(80.9)	54	(19.1)	283	94.6
Actual negative outcome	Rational	5	False positives (31.3)	11	Hits (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)				

Table 3. Categorical outcomes by signal-alarm and non-signal-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%)

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,  $\chi^2(1, n = 299) = 14.18, p < 0.0001$ , implying that the rational method was often accurate in differentiating between clients whose outcomes are successful and clients who did not show positive change in therapy. Among clients falsely predicted to fail by the rational method, 56% showed no reliable change while 44% made reliable positive change. Using deterioration as the negative outcome criterion, the difference in hit rates is virtually identical

to that when using reliable worsening as the negative outcome criterion, with a 79.9% overall hit rate for the empirical method versus a 61.2% hit rate for the rational method.

#### Accuracy of Predictive Models: Continuous Outcomes

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 very small ES of 0.17, which is slightly less than the ES of 0.20 widely considered to represent a small effect size (Cohen, 1988). This indicates that little improvement, on average, occurred for those clients labeled as signal-alarms 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 = 0.53), indicating that the average outcome for a client predicted to fail by the rational method was actually substantially positive. This also shows a notable contrast to clients predicted to fail by the empirical method, who only improved by a mean ES of 0.17. Clients predicted to have a non-negative outcome (i.e. *not* to have a negative response to treatment) by either method

showed similarly positive outcomes (ES for positive prediction by empirical method = 0.90; ES for positive prediction by rational method = 0.88). Table 4 summarizes the above results.

### Accuracy of Subcategories: Categorical Outcomes

The rational method's red category caught nine of 16 clients who showed reliable worsening, whereas its yellow category identified two clients who worsened. The empirical method's red category identified 12 of 16 clients who worsened and its yellow method detected one client who became reliably worse over the course of treatment.

Categorical outcomes are provided for each method's color-coded prediction subcategories in Table 5. For the empirical method, each progressively more positive prediction subcategory was related to a smaller proportion of cases who became reliably worse and a higher percentage of cases who reliably improved. Contrary to our hypothesis, the rational method's subcategories did not predict progressively more positive outcomes (less worsening and more reliable improvement) as the subcategories became more positive in their predictions.

Table 4. Change in OQ-45 scores by rational or empirical prediction of outcome

Method	Prediction	Mean ES change	Mean OQ-45 change
Rational	Negative	0.53	12.03
Empirical	Negative	0.17	3.79
Rational	Positive	0.88	19.9
Empirical	Positive	0.90	20.39

Table 5. Outcomes by prediction subcategories

Method	Category	Reliably worse		Reliably improved		No reliable change		ES change	Total	%
		N	%	N	%	N	%			
Rational	Red	9	(15.3)	18	(30.5)	32	(54.2)	0.36	59	19.7
	Yellow	2	(3.1)	31	(48.4)	31	(48.4)	0.69	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)	0.73	75	25.1
Empirical	Red	12	(22.2)	12	(22.2)	30	(55.6)	0.07	54	18.1
	Yellow	1	(7.7)	6	(46.2)	6	(46.2)	0.58	13	4.3
	Green	3	(1.6)	102	(54.0)	84	(44.4)	0.75	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

### Accuracy of Subcategories: Continuous Outcomes

Analyses were conducted to see how much the average client changed within each subcategory of prediction for each method. This analysis investigates the magnitude or clinical meaningfulness of change. The results are shown in Table 5. The average client in the red category of the rational method made small improvement, using Cohen's definition of a small ES (Cohen, 1988). For the rational method, those in the yellow category generally showed moderate change, and those labeled as green generally experienced notable change denoted by a large effect size, yet those labeled as most likely to succeed, clients in the white category, showed moderate change, less than was observed in the green category. This result ran contrary to the hypothesis that each increasingly optimistic prediction category would yield greater positive average outcomes.

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, whereas each increasingly optimistic prediction was related to an increased average effect of treatment. Examining both continuous and categorical results, the actual outcomes of the clients in this sample were more consistent with the predictions made by the empirical method than with those made by the rational method.

### Speed of Identification

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 ( $z = 3.83$ ,  $p < 0.0001$ ) and suggests that the rational method was quicker to issue predictive alarms for cases anticipated to have negative outcome. Of these cases more quickly identified by the rational method, four deteriorated (21.1%), eight made reliable improvement (42.1%) and seven (36.8%) made no reliable change.

### *Sensitivity at Varying Levels of Initial Distress*

At the higher end of intake OQ-45 scores (76 or greater), the rational method tended to predict negative outcomes at a much higher rate than did the empirical method. For this group of clients, the rational method generated 91 signal-alarms compared with 28 by the empirical method, a significant difference,  $\chi^2(1, n = 177) = 50.24$ ,  $p < 0.0001$ . Differences at lower levels of initial distress were not significant.

## DISCUSSION

In this investigation, the empirical method emerged as superior to the rational method in overall predictive accuracy. However, unlike the prior comparison of the two methods (Lambert et al., 2002a), there was no significant difference in the accurate identification of cases that had a negative outcome; rather, the empirical method was much more accurate in forecasting positive outcomes. The rational method generated a high percentage of false signal-alarms. In the current study, 40% of clients who showed a non-negative outcome were predicted to fail by the rational method, whereas the empirical method had a false negative rate of 19%. The percentage of false negatives for the empirical method was consistent with the prior comparison of the two methods (Lambert et al., 2002a), while the rate of false negatives for the rational method was nearly twice as high as in the previous study.

It is not clear why the rational method so frequently wrongly predicted negative outcomes. Lambert et al. (2002a) 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. Not only is this unnecessary, but changing from an effective therapy to some other intervention could potentially bring about less desirable outcomes. Actually, the provision of false negative predictions could contribute to the enactment of a negative self-fulfilling prophecy. If a system consistently generates false negative feedback, as did the rational method in the current study, then its utility may be reduced, as therapists will likely grow weary of a system that frequently makes erroneous negative predictions.

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, a system that provides excessive negative feedback to therapists runs the risk of simply being ignored. Nevertheless, it should be noted that the four feedback studies which utilized the rational method indicated significantly improved outcomes for clients who had signal-alarms, even though the base rates for alarms in those studies ranged from 18 to 40%. The base rate for signal-alarms varies widely across therapists and in the feedback studies no feedback was given in half of a therapist's cases, thus lowering the base rate in research protocols (i.e., even if the base rate is 40%, in a research protocol this would amount to an average of 20% of clients being signal alarms). Continued research on base rates of alarm signals is needed in order to fully understand the consequences of giving a false alarm.

Consistent with past research, the red alarm predicted a more negative outcome than did the yellow alarm. This was especially the case for the empirical method. The empirical red signal detected 12 of 16 clients who became reliably worse during treatment whereas the rational red

signal detected 9 of 16 cases. The rational yellow signal detected two cases who worsened reliably while the empirical yellow signal detected only one case. These findings suggest, in accordance with prior research, that the red alarm is to be taken much more seriously than a yellow alarm. Only 22 and 30% of clients receiving red alarms by the empirical and rational methods, respectively, showed reliable positive change. In the prior investigation (Lambert et al., 2002a), these numbers were lower—12 and 14%—but this study's lower data collection rate led to the generation of fewer signal-alarms, which very likely accounts for the relatively more positive outcomes of red alarm cases in the present study. However, even with a lower data collection rate, the red alarm was still generally a prognosticator of poor outcome.

Consistent with the prior study (Lambert et al., 2002a), the rational method was quicker to identify signal-alarm cases. Given that prior research has indicated that clients labeled as signal-alarms within the first three sessions were more likely to show poor final outcomes than those identified later in treatment, this may be of some significance. However, it is important to note that of the 19 cases identified as signal-alarms earlier by the rational method only four went on to show negative outcome. The rational method was designed to be especially sensitive to signs of treatment failure for clients who present as highly disturbed. Unfortunately, in the present study, it appears that the rational method was overly sensitive, as 61 of its 65 negative predictions for clients with an initial OQ-45 of 87 or higher were incorrect. While only four of these clients had negative outcomes, 36 improved reliably, a very poor display of predictive power. Clients who enter therapy with high OQ-45 scores are probably the most likely to suffer the most deleterious consequences should psychotherapy result in negative outcome. Unfortunately, the rational method was highly inaccurate for this important group of clients, limiting its usefulness among those for whom inaccurate outcome forecasting is likely most harmful.

The relatively low data collection rate (at only 49% of sessions were OQ-45 data collected) may appear initially as a major weakness of this investigation. 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. However, outside of missing three cases that reliably worsened, the empirical method had a good hit rate, comparable to a prior investigation with a much

higher data collection rate (Lambert et al., 2002a). The predictions of the empirical method were neatly related to the average effect of treatment in a linear fashion, with red cases showing the worst outcomes and blue cases, on average, doing quite well in treatment. For the rational method, clients in the 'white' group, predicted to show the most positive change, only showed a mean ES change of 0.04 greater than the average yellow-alarm client, which casts doubt on the validity of both yellow and white feedback. 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 the rational method would be adversely impacted more than the empirical method.

While the low rate of data collection is a limitation, it can also be viewed as a strength. In daily clinical practice, it is improbable that regular administration of the OQ-45, or any other outcome measure, will occur in most settings. Secretarial personnel are often in charge of data collection and they, of course, often perform other tasks in addition to administering outcome measures. In fact, delivering outcome measures to clients may be viewed by many secretaries as an aversive task (e.g., asking people to do one more thing that they may not want to do can be difficult), leading them to avoid performing it. Further, when a rush of clients arrives at the top of an hour, it may often be difficult to ensure that each client completes an OQ-45 prior to the session. Clients sometimes arrive late to session, in which case therapists often feel pressured to spend as much productive time as possible in session, not wanting to yield another five to ten minutes of valuable therapy time to data collection. Thus, the results gathered in this study may indeed be more applicable to general clinical practice settings than those generated from research in which a very high rate of OQ-45 administration occurred.

Because 72% 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%) 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 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, potentially less pathological sample in this study introduces a problem of restricted range. It is likely that a comparison of these two predictive 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 variables. The prior statement is speculative and should be investigated through further research comparing the two methods in a more treatment resistant population, such as a community mental health center.

The conclusions that can be drawn from this study are limited by the definitions of outcome that were utilized. While the use of reliable positive and negative change has precedent in the literature (Jacobson & Truax, 1991), there are certainly other methods of classifying clinically meaningful change following therapy. The use of effect size change in this study was a supplement to the use of RCI as an outcome metric. The use of other, perhaps more comprehensive, methods and measures would be a welcome addition. However, it is certainly difficult to expect clients to undergo repeated assessment with more thorough and time-consuming measures. Indeed, we argue that the use of lengthier repeated assessments may be useful in research settings but is highly unlikely to occur in daily clinical practice settings.

The present study supports the hypothesis that the empirical method more accurately predicts psychotherapy outcomes than the rational method. In the previous investigation the rational method generally underperformed relative to the empirical method, but in this study the difference between methods was of a larger magnitude, particularly regarding the generation of false negative feedback. The prior comparison of rational and empirical methods (Lambert et al., 2002a) found a significant advantage for the empirical method in identifying negative outcomes, and the current study found that the empirical method generated significantly fewer false alarms while providing, as opposed to the rational method, a more positive outcome for the average client with each progressively more positive level of prediction. These results strongly suggest that the empirical method is to be preferred in future studies and in clinical practice where feedback is provided to therapists based on client progress.

It may also be useful to develop empirical predictive models with different cut-offs than the current model. This could be useful in identifying clients who are unlikely to show a positive treat-

ment response as opposed to those predicted to show a negative response. Feedback research could then be done to see whether those predicted to show little response show a greater response due to therapist notification of the likelihood of nonresponse and subsequent alteration of treatment.

Based on two investigations with a total of 791 clients, it appears that the empirical method is more accurate than the rational method in forecasting psychotherapy outcomes. Feedback based on the rational method was effective in enhancing outcomes in four prior studies, and it stands to reason that using the empirical method should result in even greater gains for clients whose therapists receive feedback, because empirically generated predictions are of greater predictive validity than predictions generated by the rational method.

## REFERENCES

- 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.
- Bauer, S., Lambert, M.J., & Nielsen, S.L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82*, 60–70.
- 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.
- Beckstead, D.J., Hatch, A.L., Lambert, M.J., Eggett, D.L., Goates, M.K., & Vermeersch, D.A. (2003). Clinical significance of the Outcome Questionnaire (OQ-45.2). *The Behavior Analyst Today, 4*, 86–97.
- Breslin, F., Sobell, L.C., Buchan, G., & Cunningham, J. (1997). Toward a stepped-care approach to treating problem drinkers: The predictive validity of within-treatment variables and therapist prognostic ratings. *Addiction, 92*, 1479–1489.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- 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.
- 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.
- Friedman, P.H. (1994). *Friedman well-being scale*. Redwood City, CA: Mind Garden.
- Garb, H.N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*(3), 387–396.

- Gaynor, S.T., Weersing, V.R., Kolko, D.J., Birmaher, B., Heo, J., & Brent, D.A. (2003). The prevalence and impact of large sudden improvements during adolescent therapy for depression: A comparison across cognitive-behavioral, family, and supportive therapy. *Journal of Consulting and Clinical Psychology, 71*(2), 386–393.
- Haas, E., Hill, R.D., Lambert, M.J., & Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology, 58*(9), 1157–1172.
- Hannan, C., Lambert, M.J., Harmon, C., Nielsen, S.L., Smart, D.M., Shimokawa, K., & Sutton, S.W. (2005). A lab test and algorithms for identifying patients at risk for treatment failure. *Journal of Clinical Psychology: In Session, 61*(2), 155–163.
- Hannover, W., Richard, M., Hansen, N.B., Martinovich, Z., & Kordy, H. (2002). A classification tree model for decision-making in clinical practice: An application based on the data of the German multicenter study on eating disorders, project TR-EAT. *Psychotherapy Research, 12*(4), 445–461.
- Hawkins, E.J., Lambert, M.J., Vermeersch, D.A., Slade, K., & Tuttle, K. (2004). The therapeutic effects of providing client progress information to patients and therapists. *Psychotherapy Research, 10*, 308–327.
- Howard, K.I., Kopta, S.M., Krause, M.S., & Orlinsky, D.E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159–164.
- Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.
- Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart-Heidelberg Model. *Journal of Consulting and Clinical Psychology, 69*(2), 173–183.
- Lambert, M.J., Hansen, N.B., & Finch, A.E. (2001a). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*(2), 159–172.
- Lambert, M.J., Huefner, J.C., & Reisinger, C.W. (2000). Quality improvement: Current research in outcome management. In G. Stricker, W.G. Trow, & S.A. Shueman (Eds.), *Handbook of quality management in behavioral health* (pp. 95–110). New York: Kluwer-Plenum.
- Lambert, M.J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R.C., Shimokawa, K., Christopher, C., & Burlingame, G.M. (2004). *Administration and Scoring Manual for the Outcome Questionnaire-45*. Orem, UT: American Professional Credentialing Services.
- Lambert, M.J., & Ogles, B.M. (2004). The efficacy and effectiveness of psychotherapy. In M.J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193). New York: Wiley.
- Lambert, M.J., Whipple, J.L., Bishop, M.J., & Vermeersch, D.A. (2002a). 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. (2001b). 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. (2002b). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9*(2), 91–103.
- 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.
- Lunnen, K.M., & Ogles, B.A. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 400–410.
- Mohr, D.C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice, 2*(1), 1–27.
- 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.
- Richard, M., & Kordy, H. (2000, June). Early treatment response: Conceptualization, predictive validity and application in quality management. Paper presented at the annual meetings of the Society for Psychotherapy Research, Chicago, IL.
- Smith, M., Glass, G., & Miller, T. (1980). *The benefits of psychotherapy*. Baltimore, MD: The Johns Hopkins University Press.
- Speer, D.C., & Greenbaum, P.E. (1995). Five methods for computing significant individual patient change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63*, 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.
- Stiles, W.B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D.A. et al. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology, 71*(1), 14–21.
- Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*(4), 522–532.
- Tang, T.A., & DeRubies, R.J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 67*, 894–904.
- Vermeersch, D.A., Lambert, M.J., & Burlingame, G.M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*(2), 242–261.
- Ware, J., Kosinski, M., & Keller, S.D. (1994). *SF-36 Physical and Mental Health Summary Scales: A user's manual*. Boston, MA: The Health Institute, New England Medical Center.

- Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69*(6), 875–899.
- 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.
- Wilson, G.T. (1999). Rapid response to cognitive behavior therapy. *Clinical Psychology: Science and Practice, 6*, 289–292.
- Zung, W.W. (1971). A rating instrument for anxiety disorders. *Psychosomatics, 6*, 371–379.