The Effectiveness of Psychotherapy

The Consumer Reports *Study*

Martin E. P. Seligman
University of Pennsylvania

Consumer Reports *(1995, November) published an article which concluded that patients benefited very substantially from psychotherapy, that long-term treatment did considerably better than short-term treatment, and that psychotherapy alone did not differ in effectiveness from medication plus psychotherapy. Furthermore, no specific modality of psychotherapy did better than any other for any disorder; psychologists, psychiatrists, and social workers did not differ in their effectiveness as treaters; and all did better than marriage counselors and long-term family doctoring. Patients whose length of therapy or choice of therapist was limited by insurance or managed care did worse. The methodological virtues and drawbacks of this large-scale survey are examined and contrasted with the more traditional efficacy study, in which patients are randomized into a manualized, fixed duration treatment or into control groups. I conclude that the* Consumer Reports *survey complements the efficacy method, and that the best features of these two methods can be combined into a more ideal method that will best provide empirical validation of psychotherapy.*

How do we find out whether psychotherapy works? To answer this, two methods have arisen: the *efficacy study* and the *effectiveness study*. An efficacy study is the more popular method. It contrasts some kind of therapy to a comparison group under well-controlled conditions. But there is much more to an efficacy study than just a control group, and such studies have become a high-paradigm endeavor with sophisticated methodology. In the ideal efficacy study, all of the following niceties are found:

1. The patients are randomly assigned to treatment and control conditions.
2. The controls are rigorous: Not only are patients included who receive no treatment at all, but placebos containing potentially therapeutic ingredients credible to both the patient and the therapist are used in order to control for such influences as rapport, expectation of gain, and sympathetic attention (dubbed *nonspecifics*).
3. The treatments are manualized, with highly detailed scripting of therapy made explicit. Fidelity to the manual is assessed using videotaped sessions, and wayward implementers are corrected.
4. Patients are seen for a fixed number of sessions.
5. The target outcomes are well operationalized (e.g., clinician-diagnosed DSM–IV disorder, number of reported orgasms, self-reports of panic attacks, percentage of fluent utterances).
6. Raters and diagnosticians are blind to which group the patient comes from. (Contrary to the "double-blind" method of drug studies, efficacy studies of psychotherapy can be at most "single-blind," since the patient and therapist both know what the treatment is. Whenever you hear someone demanding the double-blind study of psychotherapy, hold onto your wallet.)
7. The patients meet criteria for a single diagnosed disorder, and patients with multiple disorders are typically excluded.
8. The patients are followed for a fixed period after termination of treatment with a thorough assessment battery.

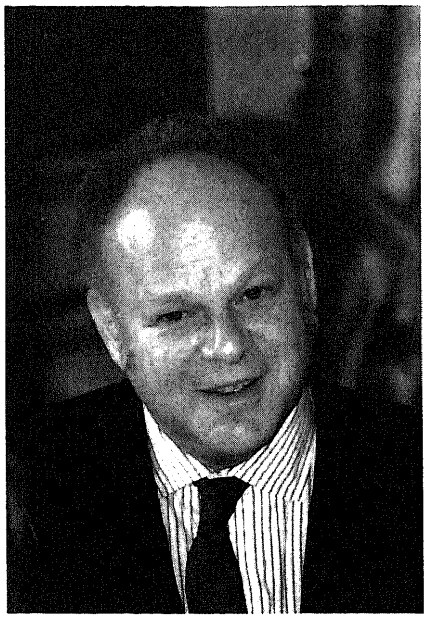
So when an efficacy study demonstrates a difference between a form of psychotherapy and controls, academic clinicians and researchers take this modality seriously indeed. In spite of how expensive and time-consuming they are, hundreds of efficacy studies of both psychotherapy and drugs now exist—many of them well done. These studies show, among many other things, that cognitive therapy, interpersonal therapy, and medications all provide moderate relief from unipolar depressive disorder; that exposure and clomipramine both relieve the symptoms of obsessive–compulsive disorder moderately well but that exposure has more

Editor's note. Gary R. VandenBos served as action editor for this article.

Author's note. I thank the staff of *Consumer Reports* for their cooperation and generosity. The opinions expressed in this article are solely my own and do not represent the opinions of *Consumer Reports*. Among the many people at *Consumer Reports* who contributed to this project, I want to single out Mark Kotkin, Joel Gurin, Donato Vaccaro, and Rochelle Green. Mark was unflagging in his probing of the data and in his appetite for brainstorming at any hour of the day or night. The project was Joel's brainchild and he shepherded it through from beginning to end. Donato provided a bridge between the *Consumer Reports* perspective and the mental health professional's perspective. Rochelle unblinkingly reported the findings to the readers. I also thank Neil Jacobson, Ken Howard, Lee Sechrest, David Seligman, Timothy Stickle, Michelle Stewart-Fouts, and George Stricker for comments on various issues raised by this data set. PHS Grant MH19604 partially supported the writing of this article.

Correspondence concerning this article should be addressed to Martin E. P. Seligman, Department of Psychology, 3815 Walnut Street, University of Pennsylvania, Philadelphia, PA 19104.

 0003-066X/95/$2.00
Vol. 50, No. 12, 965–974



Martin E. P. Seligman

lasting benefits; that cognitive therapy works very well in panic disorder; that systematic desensitization relieves specific phobias; that "applied tension" virtually cures blood and injury phobia; that transcendental meditation relieves anxiety; that aversion therapy produces only marginal improvement with sexual offenders; that disulfram (Antabuse) does not provide lasting relief from alcoholism; that flooding plus medication does better in the treatment of agoraphobia than either alone; and that cognitive therapy provides significant relief of bulimia, outperforming medications alone (see Seligman, 1994, for a review).

The high praise "empirically validated" is now virtually synonymous with positive results in efficacy studies, and many investigators have come to think that an efficacy study is the "gold standard" for measuring whether a treatment works.

I also had come to that opinion when I wrote *What You Can Change & What You Can't* (Seligman, 1994). In trying to summarize what was known about the effects of the panoply of drugs and psychotherapies for each major disorder, I read hundreds of efficacy studies and came to appreciate the genre. At minimum I was convinced that an efficacy study may be the best scientific instrument for telling us whether a novel treatment is *likely* to work on a given disorder when the treatment is exported from controlled conditions into the field. Because treatment in efficacy studies is delivered under tightly controlled conditions to carefully screened patients, sensitivity is maximized and efficacy studies are very useful for deciding whether one treatment is better than another treatment for a given disorder.

But my belief has changed about what counts as a "gold standard." And it was a study by *Consumer Reports* (1995, November) that singlehandedly shook my belief. I came to see that deciding whether one treatment, under highly controlled conditions, works better than another treatment or a control group is a different question from deciding what works in the field (Muñoz, Hollon, McGrath, Rehm, & VandenBos, 1994). I no longer believe that efficacy studies are the only, or even the best, way of finding out what treatments actually work in the field. I have come to believe that the "effectiveness" study of how patients fare under the actual conditions of treatment in the field, can yield useful and credible "empirical validation" of psychotherapy and medication. This is the method that *Consumer Reports* pioneered.

What Efficacy Studies Leave Out

It is easy to assume that, if some form of treatment is not listed among the many which have been "empirically validated," the treatment must be inert, rather than just "untested" given the existing method of validation. I will dub this the *inertness assumption.* The inertness assumption is a challenge to practitioners, since long-term dynamic treatment, family therapy, and more generally, eclectic psychotherapy, are not on the list of treatments empirically validated by efficacy studies, and these modalities probably make up most of what is actually practiced. I want to look closely at the inertness assumption, since the effectiveness strategy of empirical validation follows from what is wrong with the assumption.

The usual argument against the inertness assumption is that long-term dynamic therapy, family therapy, and eclectic therapy cannot be tested in efficacy studies, and thus we have no hard evidence one way or another. They cannot be tested because they are too cumbersome for the efficacy study paradigm. Imagine, for example, what a decent efficacy study of long-term dynamic therapy would require: control groups receiving no treatment for several years; an equally credible comparison treatment of the same duration that has the same "nonspecifics"—rapport, attention, and expectation of gain—but is actually inert; a step-by-step manual covering hundreds of sessions; and the random assignment of patients to treatments which last a year or more. The ethical and scientific problems of such research are daunting, to say nothing of how much such a study would cost.

While this argument cannot be gainsaid, it still leaves the average psychotherapist in an uncomfortable position, with a substantial body of literature validating a panoply of short-term therapies the psychotherapist does not perform, and with the long-term, eclectic therapy he or she does perform unproven.

But there is a much better argument against the inertness assumption: *The efficacy study is the wrong method for empirically validating psychotherapy as it is actually done, because it omits too many crucial elements of what is done in the field.*

The five properties that follow characterize psychotherapy as it is done in the field. Each of these properties are absent from an efficacy study done under controlled conditions. If these properties are important to patients' getting better, efficacy studies will underestimate or even miss altogether the value of psychotherapy done in the field.

1. Psychotherapy (like other health treatments) in the

field is *not of fixed duration.* It usually keeps going until the patient is markedly improved or until he or she quits. In contrast, the intervention in efficacy studies stops after a limited number of sessions—usually about 12—regardless of how well or how poorly the patient is doing.

2. Psychotherapy (again, like other health treatments) in the field is *self-correcting.* If one technique is not working, another technique—or even another modality—is usually tried. In contrast, the intervention in efficacy studies is confined to a small number of techniques, all within one modality and manualized to be delivered in a fixed order.

3. Patients in psychotherapy in the field often get there by *active* shopping, entering a kind of treatment they actively sought with a therapist they screened and chose. This is especially true of patients who work with independent practitioners, and somewhat less so of patients who go to outpatient clinics or have managed care. In contrast, patients enter efficacy studies by the *passive* process of random assignment to treatment and acquiescence with who and what happens to be offered in the study (Howard, Orlinsky, & Lueger, 1994).

4. Patients in psychotherapy in the field usually have *multiple problems,* and psychotherapy is geared to relieving parallel and interacting difficulties. Patients in efficacy studies are selected to have but one diagnosis (except when two conditions are highly comorbid) by a long set of exclusion and inclusion criteria.

5. Psychotherapy in the field is almost always concerned with *improvement in the general functioning* of patients, as well as amelioration of a disorder and relief of specific, presenting symptoms. Efficacy studies usually focus only on specific symptom reduction and whether the disorder ends.

It is hard to imagine how one could ever do a scientifically compelling efficacy study of a treatment which had variable duration and self-correcting improvisations and was aimed at improved quality of life as well as symptom relief, with patients who were not randomly assigned and had multiple problems. But this does not mean that the effectiveness of treatment so delivered cannot be empirically validated. Indeed it can, but it requires a different method: a survey of large numbers of people who have gone through such treatments. So let us explore the virtues and drawbacks of a well-done effectiveness study, the *Consumer Reports* (1995) one, in contrast to an efficacy study.

Consumer Reports *Survey*

Consumer Reports (*CR*) included a supplementary survey about psychotherapy and drugs in one version of its 1994 annual questionnaire, along with its customary inquiries about appliances and services. *CR*'s 180,000 readers received this version, which included approximately 100 questions about automobiles and about mental health. *CR* asked readers to fill out the mental health section "if at any time over the past three years you experienced stress or other emotional problems for which you sought help from any of the following: friends, relatives, or a member of the clergy; a mental health professional like a psychologist or a psychiatrist; your family doctor; or a support group." Twenty-two thousand readers responded. Of these, approximately 7,000 subscribers responded to the mental health questions. Of these 7,000, about 3,000 had just talked to friends, relatives, or clergy, and 4,100 went to some combination of mental health professionals, family doctors, and support groups. Of these 4,100, 2,900 saw a mental health professional: Psychologists (37%) were the most frequently seen mental health professional, followed by psychiatrists (22%), social workers (14%), and marriage counselors (9%). Other mental health professionals made up 18%. In addition, 1,300 joined self-help groups, and about 1,000 saw family physicians. The respondents as a whole were highly educated, predominantly middle class; about half were women, and the median age was 46.

Twenty-six questions were asked about mental health professionals, and parallel but less detailed questions were asked about physicians, medications, and self-help groups:

- What kind of therapist
- What presenting problem (e.g., general anxiety, panic, phobia, depression, low mood, alcohol or drugs, grief, weight, eating disorders, marital or sexual problems, children or family, work, stress)
- Emotional state at outset (from *very poor* to *very good*)
- Emotional state now (from *very poor* to *very good*)
- Group versus individual therapy
- Duration and frequency of therapy
- Modality (psychodynamic, behavioral, cognitive, feminist)
- Cost
- Health care plan and limitations on coverage
- Therapist competence
- How much therapy helped (from *made things a lot better* to *made things a lot worse*) and in what areas (specific problem that led to therapy, relations to others, productivity, coping with stress, enjoying life more, growth and insight, self-esteem and confidence, raising low mood)
- Satisfaction with therapy
- Reasons for termination (problems resolved or more manageable, felt further treatment wouldn't help, therapist recommended termination, a new therapist, concerns about therapist's competence, cost, and problems with insurance coverage)

The data set is thus a rich one, probably uniquely rich, and the data analysis was sophisticated. Because I was privileged to be a consultant to this study and thus privy to the entire data set, much of what I now present will be new to you—even if you have read the *CR* article carefully. *CR*'s analysts decided that no single measure of therapy effectiveness would do and so created a multivariate measure. This composite had three subscales, consisting of:

1. Specific improvement ("How much did treatment help with the specific problem that led you to therapy?"

made no difference; *made things somewhat worse*; *made things a lot worse*; *not sure*);

2. Satisfaction ("Overall how satisfied were you with this therapist's treatment of your problems?" *completely satisfied*; *very satisfied*; *fairly well satisfied*; *somewhat satisfied*; *very dissatisfied*; *completely dissatisfied*); and

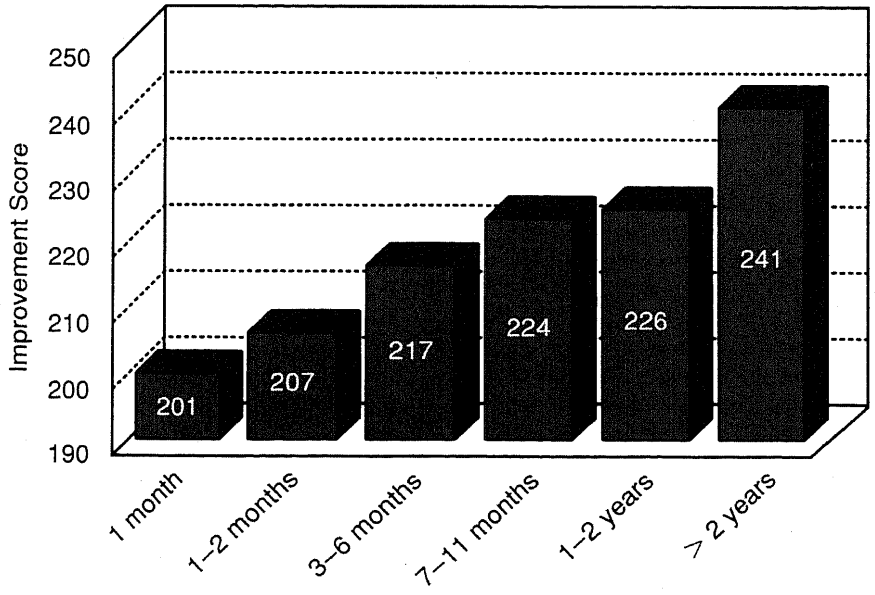
3. Global improvement (how respondents described their "overall emotional state" at the time of the survey compared with the start of treatment: "*very poor*: I barely managed to deal with things; *fairly poor*: Life was usually pretty tough for me; *so-so*: I had my ups and downs; *quite good*: I had no serious complaints; *very good*: Life was much the way I liked it to be").

Each of the three subscales was transformed and weighted equally on a 0–100 scale, resulting in a 0–300 scale for effectiveness. The statistical analysis was largely multiple regression, with initial severity and duration of treatment (the two biggest effects) partialed out. Stringent levels of statistical significance were used.

There were a number of clear-cut results, among them:

- Treatment by a mental health professional usually worked. Most respondents got a lot better. Averaged over all mental health professionals, of the 426 people who were feeling *very poor* when they began therapy, 87% were feeling *very good*, *good*, or at least *so-so* by the time of the survey. Of the 786 people who were feeling *fairly poor* at the outset, 92% were feeling *very good*, *good*, or at least *so-so* by the time of the survey. These findings converge with meta-analyses of efficacy (Lipsey & Wilson, 1993; Shapiro & Shapiro, 1982; Smith, Miller, & Glass, 1980).
- Long-term therapy produced more improvement than short-term therapy. This result was very robust, and held up over all statistical models. Figure 1 plots the overall rating (on the 0–300 scale defined above) of improvement as a function of length of treatment. This "dose–response curve" held for patients in both psychotherapy alone and in psychotherapy plus medication (see Howard, Kopta, Krause, & Orlinsky, 1986, for parallel dose–response findings for psychotherapy).
- There was no difference between psychotherapy alone and psychotherapy plus medication for any disorder (very few respondents reported that they had medication with no psychotherapy at all).
- While all mental health professionals appeared to help their patients, psychologists, psychiatrists, and social workers did equally well and better than marriage counselors. Their patients' overall improvement scores (0–300 scale) were 220, 226, 225 (not significantly different from each other), and 208 (significantly worse than the first three), respectively.
- Family doctors did just as well as mental health professionals in the short term, but worse in the long term. Some patients saw both family doctors and

Figure 1
Duration of Therapy



Note. $N = 2,846$. The 300-point scale is derived from the unweighted sum of responses to three 100-point subscales. The subscales measured specific improvement (i.e., how much treatment helped with problems that led to therapy), satisfaction with therapist, and global improvement (i.e., how respondents felt at time of survey, compared with when they began treatment).

mental health professionals, and those who saw both had more severe problems. For patients who relied solely on family doctors, their overall improvement scores when treated for up to six months was 213, and it remained at that level (212) for those treated longer than six months. In contrast, the overall improvement scores for patients of mental health professionals was 211 up to six months, but climbed to 232 when treatment went on for more than six months. The advantages of long-term treatment by a mental health professional held not only for the specific problems that led to treatment, but for a variety of general functioning scores as well: ability to relate to others, coping with everyday stress, enjoying life more, personal growth and understanding, self-esteem and confidence.

- Alcoholics Anonymous (AA) did especially well, with an average improvement score of 251, significantly bettering mental health professionals. People who went to non-AA groups had less severe problems and did not do as well as those who went to AA (average score = 215).
- Active shoppers and active clients did better in treatment than passive recipients (determined by responses to "Was it mostly your idea to seek therapy? When choosing this therapist, did you discuss qualifications, therapist's experience, discuss frequency, duration, and cost, speak to someone who was treated by this therapist, check out other therapists? During therapy, did you try to be as open as possible, ask for explanation of diagnosis and unclear terms, do homework, not cancel sessions often, discuss negative feelings toward therapist?").
- No specific modality of psychotherapy did any better than any other for any problem. These results confirm the "dodo bird" hypothesis, that all forms of psychotherapies do about equally well (Luborsky, Singer, & Luborsky, 1975). They come as a rude shock to efficacy researchers, since the main theme of efficacy studies has been the demonstration of the usefulness of specific techniques for specific disorders.
- Respondents whose choice of therapist or duration of care was limited by their insurance coverage did worse, as presented in Table 1 (determined by responses to "Did limitations on your insurance coverage affect any of the following choices you made? Type of therapist I chose; How often I met with my therapist; How long I stayed in therapy").

These findings are obviously important, and some of them could not be included in the original *CR* article because of space limitations. Some of these findings were quite contrary to what I expected, but it is not my intention to discuss their substance here. Rather, I want to explore the methodological adequacy of this survey. My underlying questions are "Should we believe the findings?" and "Can the method be improved to give more authoritative answers?"

Consumer Reports *Survey: Methodological Virtues*

Sampling. This survey is, as far as I have been able to determine, the most extensive study of psychotherapy effectiveness on record. The sample is not representative of the United States as a whole, but my guess is that it is roughly representative of the middle class and educated population who make up the bulk of psychotherapy patients. It is important that the sample represents people who choose to go to treatment for their problems, not people who do not "believe in" psychotherapy or drugs. The *CR* sample, moreover, is probably weighted toward "problem solvers," people who actively try to do something about what troubles them.

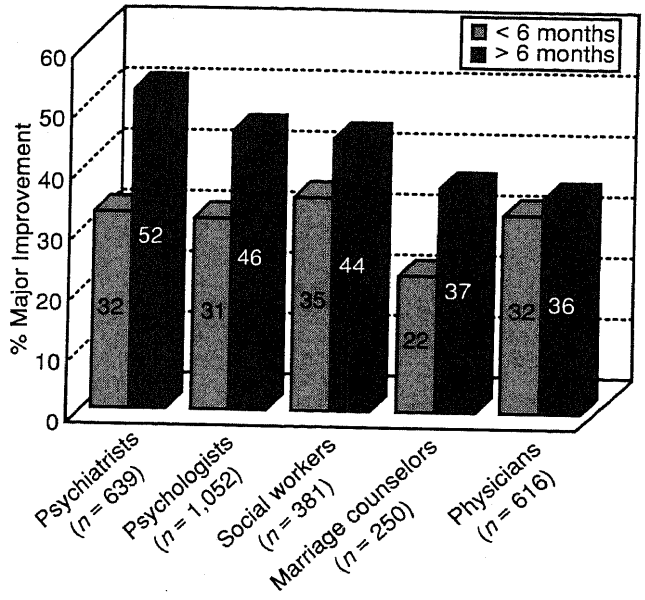
Treatment duration. *CR* sampled all treatment durations from one month or less through two years or more. Because the study was naturalistic, treatment, it can be supposed, continued until the patient (a) was better, (b) gave up unimproved, or (c) had his or her coverage run out. This, by definition, mirrors what actually happens in the field. In contrast to all efficacy studies, which are of fixed treatment duration regardless of how the patient is progressing, the *CR* study informs us about treatment effectiveness under the duration constraints of actual therapy.

Table 1
Limitations on Insurance Coverage and Improvement

Limitations on your insurance coverage	Percent checking item[a]	Coverage limited		Coverage not limited	
		Overall score	Specific improvement	Overall score	Specific improvement
Type of therapist I chose	20	211	77	224	83
How often I met with my therapist	26	214	79	224	82
How long I stayed in therapy	24	212	78	224	83
Percent of any of the above	43	212	78	226	83

Note. $N = 2{,}900$. All differences for the overall scores were statistically significant at $p < .01$. The same held true for the specific score, except for "How often I met with my therapist," which was significant at $p < .05$. Statistical controls for both severity and duration were applied. Source: *Consumer Reports* 1994 Annual Questionnaire.
[a]multiple responses permitted.

Figure 2
Improvement for Presenting Symptoms



Note. $N = 2{,}738$. Percentage of respondents who reported that treatment "made things a lot better" with respect to the specific problem that led to treatment by psychiatrists, psychologists, social workers, marriage counselors, or family doctors, segregated by those treated for more than six months and those treated for less than six months.

Self-correction. Because the *CR* study was naturalistic, it informs us of how treatment works as it is actually performed—without manuals and with self-correction when a technique falters. This also contrasts favorably to efficacy studies, which are manualized and not self-correcting when a given technique or modality fails.

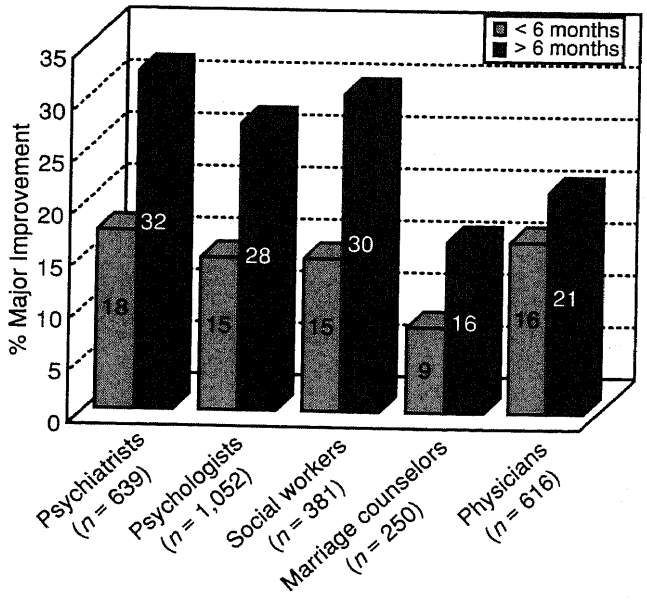
Multiple problems. The large majority of respondents in the *CR* study had more than one problem. We can also assume that a good-sized fraction were "subclinical" in their problems and would not meet *DSM–IV* criteria for any disorder. No patients were discarded because they failed exclusion criteria or because they fell one symptom short of a full-blown "disorder." Thus the sample more closely reflected people who actually seek treatment than the filtered and single-disordered patients of efficacy studies.

General functioning. The *CR* study measured self-reported changes in productivity at work, interpersonal relations, well-being, insight, and growth, in addition to improvement on the presenting problem. Improvement on the presenting problem is shown in Figure 2; improvement over work and social domains is shown in Figure 3; and improvement over personal domains is shown in Figure 4. Importantly, more improvement on the presenting problem occurred for treatments which lasted longer than six months. In addition, more improvement occurred in work, interpersonal relations, enjoyment of life, and personal growth domains in treatments which lasted longer than six months. Since improvement in general functioning, as well as symptom relief, is almost always a goal of actual treatment but rarely of efficacy studies, the *CR* study adds to our knowledge of how treatment does beyond the mere elimination of symptoms.

Clinical significance. There has been much debate about how to measure the "clinical significance" of a treatment. Efficacy studies are designed to detect statistically significant differences between a treatment and control groups, and an "effect size" can be computed. But what degree of statistical significance is clinical significance? How large an effect size is meaningful? The *CR* study leaves little doubt about the human significance of its findings, since respondents answered directly about how much therapy helped the problem that led them to treatment—from *made things a lot better* to *made things a lot worse.* Of those who started out feeling *very poor,* 54% answered treatment *made things a lot better,* and another one third answered it made things *somewhat better.*

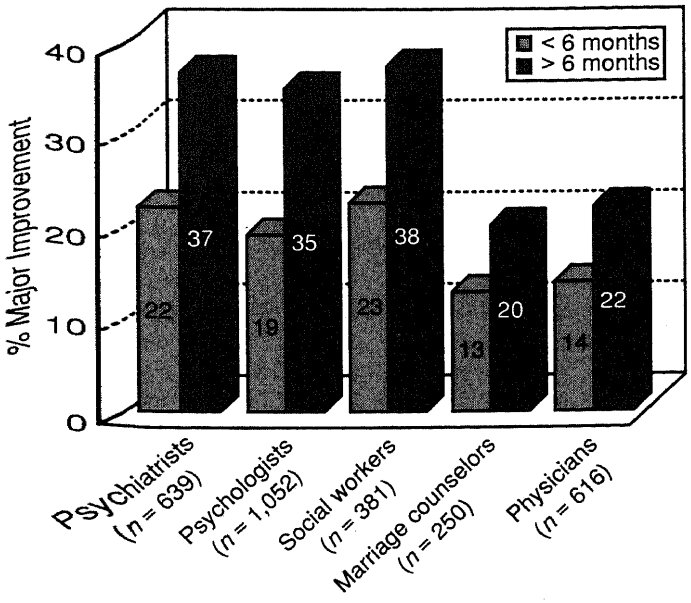
Unbiased. Finally, it cannot be ignored that *CR* is about as unbiased a scrutinizer of goods and services as exists in the public domain. They have no axe to grind for or against medications, psychotherapy, managed care, insurance companies, family doctors, AA, or long-term treatment. They do not care if psychologists do better or worse than psychiatrists, marriage and family counselors, or social workers. They are not pursuing government grants or drug company favors. They do not accept advertisements. They have a track record of loyalty only to consumers. So this study

Figure 3
Improvement Over Work and Social Domains



Note. $N = 2{,}738$. Mean percentage who reported that treatment "made things a lot better" with respect to three domains: ability to relate to others, productivity at work, and coping with everyday stress. Those treated by psychiatrists, psychologists, social workers, marriage counselors, and physicians are segregated by treatment for more than six months versus treatment for less than six months.

Figure 4
Improvement Over Personal Domains



Note. $N = 2,738$. Mean percentage who reported that treatment "made things a lot better" with respect to four domains: enjoying life more, personal growth and insight, self-esteem and confidence, and alleviating low moods. Those treated by psychiatrists, psychologists, social workers, marriage counselors, and physicians are segregated by treatment for more than six months versus treatment for less than six months.

comes with higher credibility than studies that issue from drug houses, from either APA, from consensus conferences of the National Institute of Mental Health, or even from the halls of academe.

In summary, the main methodological virtue of the *CR* study is its realism: It assessed the effectiveness of psychotherapy as it is actually performed in the field with the population that actually seeks it, and it is the most extensive, carefully done study to do this. This virtue is akin to the virtues of naturalistic studies using sophisticated correlational methods, in contrast to well-controlled, experimental studies. But because it is not a well-controlled, experimental study like an efficacy study, the *CR* study has a number of serious methodological flaws. Let us examine each of these flaws and ask to what extent they compromise the *CR* conclusions.

Consumer Reports *Study: Methodological Flaws and Rebuttals*

Sampling. Is there a bias such that those respondents who succeed in treatment selectively return their questionnaires? *CR*, not surprisingly, has gone to considerable lengths to find out if its reader's surveys have sampling bias. The annual questionnaires are lengthy and can run to 100 questions or more. Moreover, the respondents not only devote a good deal of their own time to filling these out but also pay their own postage and are not compensated. So the return rate is rather low absolutely, although the 13% return rate for this survey was normal for the annual questionnaire. But it is still possible that respondents might differ systematically from the readership as a whole. For the mental health survey (and for their annual questionnaires generally), *CR* conducted a "validation survey," in which postage was paid and the respondent was compensated. This resulted in a return rate of 38%, as opposed to the 13% uncompensated return rate, and there were no differences between data from the two samples.

The possibility of two other kinds of sampling bias, however, is notable, particularly with respect to the remarkably good results for AA. First, since AA encourages lifetime membership, a preponderance of successes—rather than dropouts—would be more likely in the three-year time slice (e.g., "Have you had help in the last three years?"). Second, AA failures are often completely dysfunctional and thus much less likely to be reading *CR* and filling out extensive readers' surveys than, say, psychotherapy failures who were unsuccessfully treated for anxiety.

A similar kind of sampling bias, to a lesser degree, cannot be overlooked for other kinds of treatment failures. At any rate, it is quite possible that there was a large oversampling of successful AA cases and a smaller oversampling of successful treatment for problems other than alcoholism.

Could the benefits of long-term treatment be an artifact of sampling bias? Suppose that people who are doing well in treatment selectively remain in treatment, and people who are doing poorly drop out earlier. In other words, the early dropouts are mostly people who fail to improve, but later dropouts are mostly people whose problem resolves. *CR* disconfirmed this possibility empirically: Respondents reported not only when they left treatment but why, including leaving because their problem was resolved. The dropout rates due to the resolution of the problem were uniform across duration of treatment (less than one month = 60%; 1–2 months = 66%; 3–6 months = 67%, 7–11 months = 67%; 1–2 years = 67%; over two years = 68%).

A more sweeping limit on generalizability comes from the fact that the entire sample *chose* their treatment. To one degree or another, each person believed that psychotherapy and/or drugs would help him or her. To one degree or another, each person acknowledged that he or she had a problem and believed that the particular mental health professional seen and the particular modality of treatment chosen would help them. One cannot argue compellingly from this survey that treatment by a mental health professional would prove as helpful to troubled people who deny their problems and who do not believe in and do not choose treatment.

No control groups. The overall improvement rates were strikingly high across the entire spectrum of treatments and disorders in the *CR* study. The vast majority of people who were feeling *very poor* or *fairly poor* when they entered therapy made "substantial" (now feeling *fairly good* or *very good*) or "some" (now feeling *so-so*) gains. Perhaps

the best news for patients was that those with severe problems got, on average, *much* better. While this may be a ceiling effect, it is a ceiling effect with teeth. It means that if you have a patient with a severe disorder now, the chances are quite good that he or she will be much better within three years. But methodologically, such high rates of improvement are a yellow flag, cautioning us that global improvement over time alone, rather than with treatment or medication, may be the underlying mechanism.

More generally, because there are no control groups, the *CR* study cannot tell us directly whether talking to sympathetic friends or merely letting time pass would have produced just as much improvement as treatment by a mental health professional. The *CR* survey, unfortunately, did not ask those who just talked to friends and clergy to fill out detailed questionnaires about the results.

This is a serious objection, but there are internal controls which perform many of the functions of control groups. First, marriage counselors do significantly worse than psychologists, psychiatrists, and social workers, in spite of no significant differences in kind of problem, severity of problem, or duration of treatment. Marriage counselors control for many of the nonspecifics, such as therapeutic alliance, rapport, and attention, as well as for passage of time. Second, there is a dose–response curve, with more therapy yielding more improvement. The first point in the dose–response curve approximates no treatment: people who have less than one month of treatment have on average an improvement score of 201, whereas people who have over two years of treatment have an average score of 241. Third, psychotherapy does just as well as psychotherapy plus drugs for all disorders, and there is such a long history of placebo controls inferior to these drugs that one can infer that psychotherapy likely would have outperformed such controls had they been run. Fourth, family doctors do significantly worse than mental health professionals when treatment continues beyond six months. An objection might be made that since total length of time in treatment—rather than total amount of contact—is the covariate, comparing family doctors who do not see their patients weekly with mental health professionals—who see their patients once a week or more—is not fair. It is, of course, possible that if family doctors saw their patients as frequently as psychologists do, the two groups would do equally well. It was notable, however, that there were a significant number of complaints about family doctors: 22% of respondents said their doctor had not "provided emotional support"; 15% said their doctor "seemed uncomfortable discussing emotional issues"; and 18% said their doctor was "too busy to spend time talking to me." At any rate, the *CR* survey shows that long-term family doctoring for emotional problems—as it is actually performed in the field—is inferior to long-term treatment by a mental health professional as it is actually performed in the field.

It is also relevant that the patients attributed their improvement to treatment and not time (determined by responses to "How much do you feel that treatment helped you in the following areas?"), and I conclude that the benefits of treatment are very unlikely to be caused by the mere passage of time. But I also conclude that the *CR* study could be improved by control groups whose members are not treated by mental health professionals, matched for severity and kind of problem (but beware of the fact that random assignment will not occur). This would allow the Bayesian inference that psychotherapy works better than talking to friends, seeing an astrologer, or going to church to be made more confidently.

Self-report. *CR*'s mental health survey data, as for cars and appliances, are self-reported. Improvement, diagnosis, insurance coverage, even kind of therapist are not verified by external check. Patients can be wrong about any of these, and this is an undeniable flaw.

But two things can be said in response. First, the noise self-reports introduce—inaccuracy about improvement, incorrectness about the nature of their problem, even inaccuracy about what kind of a therapist they saw—may be random rather than systematic, and therefore would not necessarily bias the study toward the results found. Self-report, in principle, can be either rosier or more dire than the report of an external observer. Since most respondents are probably more emotionally invested in psychotherapy than in their automobiles, however, it will take further research to determine whether the noise introduced by self-report about therapy is random or systematic.

Second, the most important potential inaccuracy produced by self-report is inaccuracy about respondents' own emotional state before and after treatment, and inaccuracy in ratings of improvement in the specific problem, in productivity at work, and in human relationships. This is, however, an ever-present inaccuracy even with an experienced diagnostician, and the correlations between self-report and diagnosis are usually quite high (not surprising, given the common method variance). Such self-reports are the blood and guts of a clinical diagnosis. But multiple observers are always a virtue, and diagnosis by a third party would improve the survey method noticeably.

Blindness. The *CR* survey is not double-blind, or even single-blind. The respondent rates his or her own emotional state, and knows what treatment he or she had. So it is possible that respondents exaggerate the virtues or vices of their treatment to comply with or to overthrow their hypotheses about what *CR* wants to find. I find this far-fetched: If nonblindness compromised readers' surveys, *CR* would have long ago ceased publishing them, since the readers' evaluations of other products and services are always nonblind. *CR* validates its data for goods and services in two ways: against manufacturers' polls and for consistency over time. Using both methods, *CR* has been unable to detect systematic distortions in its nonblind surveys of goods and services.

Inadequate outcome measures. *CR*'s indexes of improvement were molar. Responses like *made things a lot better* to the question "How much did therapy help you with the specific problems that led you to therapy?" tap into gross processes. More molecular assessment of improvement, for example, "How often have you cried in the last two weeks?" or "How many ounces of alcohol did you have yesterday?" would increase the validity of the method. Such detail would, of course, make the survey more cumbersome.

A variant of this objection is that the outcome mea-

sures were *insensitive*. This objection looms large in light of the failure to find that any modality of therapy did better than any other modality of therapy for any disorder. Perhaps if more detailed, disorder-specific measures were used, the dodo bird hypothesis would have been disconfirmed.

A third variant of this objection is that the outcome measures were poorly normed. Questions like "How satisfied were you with this therapist's treatment of your problem? *Completely satisfied, very satisfied, fairly well satisfied, somewhat dissatisfied, very dissatisfied, completely dissatisfied*," and "How would you describe your overall emotional state? *very poor*: I barely managed to deal with things; *fairly poor*: Life was usually pretty tough for me; *so-so*: I had my ups and downs; *quite good*: I had no serious complaints; *very good*: Life was much the way I liked it to be" are seat-of-the-pants items which depend almost entirely on face validity, rather than on several generations of norming. So the conclusion that 90% of those people who started off *very poor* or *fairly poor* wound up in the *very good, fairly good,* or *so-so* categories does not guarantee that they had returned to normality in any strong psychometric sense. The addition of extensively normed questionnaires like the Beck Depression Inventory would strengthen the survey method (and make it more cumbersome).

Retrospective. The *CR* respondents reported retrospectively on their emotional states. While a one-time survey is highly cost-effective, it is necessarily retrospective. Retrospective reports are less valid than concurrent observation, although an exception is worth noting: waiting for the rosy afterglow of a newly completed therapy to dissipate, as the *CR* study does, may make for a more sober evaluation.The retrospective method does not allow for longitudinal observation of the same individuals for improvement across time. Thus the benefits of long-term psychotherapy are inferred by comparing different individuals' improvements cross-sectionally. A prospective study would allow comparison of the same individuals' improvements over time.

Retrospective observation is a flaw, but it may introduce random rather than systematic noise in the study of psychotherapy effectiveness. The distortions introduced by retrospection could go either in the rosier or more dire direction, but only further research will tell us if the distortions of retrospection are random or systematic.

It is noteworthy that *Consumer Reports* generally uses two methods. One is the laboratory test, in which, for example, a car is crashed into a wall at five miles per hour, and damage to the bumper is measured. The other is the reader's survey. These two methods parallel the efficacy study and the effectiveness study, respectively, in many ways. If retrospection was a fatal flaw, *CR* would have given up the reader's survey method long ago, since reliability of used cars and satisfaction with airlines, physicians, and insurance companies depends on retrospection. Regardless, the survey method could be markedly improved by being longitudinal, in the same way as an efficacy study. Self-report and diagnosis both could be done before and after therapy, and a thorough follow-up carried out as well. But retrospective reports of emotional states will always be with us, since even in a prospective study that begins with a diagnostic interview, the patient retrospectively reports on his or her (presumably) less troubled emotional state before the diagnosis.

Therapy junkies. Perhaps the important finding that long-term therapy does so much better than short-term therapy is an artifact of therapy "junkies," individuals so committed to therapy as a way of life that they bias the results in this direction. This is possible, but it is not an artifact. Those people who spend a long time in therapy may well be "true believers." Indeed, the long-term patients are distinct: They have more severe problems initially, are more likely to have an emotional disorder, are more likely to get medications, are more likely to see a psychiatrist, and are more likely to have psychodynamic treatment than the rest of the sample. Regardless, they are probably representative of the population served by long-term therapy. This population reports robust improvement with long-term treatment in the specific problem that got them into therapy, as well as in growth, insight, confidence, productivity at work, interpersonal relations, and enjoyment of life.

Perhaps people who had two or more years of therapy are likely still to be in therapy and thus unduly loyal to their therapist. They might then be more likely to distort in a rosy direction. This seems unlikely, since a comparison of people who had over two years of treatment and then ended therapy showed the same high improvement scores as those with over two years of treatment who were still in therapy (242 and 245, respectively).

Nonrandom assignment. The possibility of such biases could be reduced by random assignment of patients to treatment, but this would undermine the central virtue of the *CR* study—reporting on the effectiveness of psychotherapy as it is actually done in the field with those patients who actually seek it. In fact, the lack of random assignment may turn out to be the crucial ingredient in the validity of the *CR* method and a major flaw of the efficacy method. Many (but assuredly not all) of the problems that bring consumers into therapy have elements of what was called "wanhope" in the middle ages and is now called "demoralization." Choice and control by a patient, in and of itself, counteracts wanhope (Seligman, 1991).

Random assignment of patients to a modality or to a particular therapist not only undercuts the remoralizing effects of treatment but also undercuts the nonrandom decisions of therapists in choice of modality for a particular patient. Consider, for example, the finding that drugs plus psychotherapy did no better than psychotherapy alone for any disorder (schizophrenia and bipolar depression were too rare for analysis in this sample). The most obvious interpretation is that drugs are useless and do nothing over and above psychotherapy. But the lack of random assignment should prevent us from leaping to that conclusion. Assume, for the moment, that therapists are canny about who needs drugs plus psychotherapy and who can do well with psychotherapy alone. The therapists assign those patients accordingly so appropriate patients get appropriate treatment. This is just the same logic as a self-correcting trajectory of treatment, in which techniques and modalities are modified with the patient's progress. This means that drugs plus psychotherapy may actually have done pretty well after all—

but only in a cannily selected subset of people.

The upshot of this is that random assignment, the prettiest of the methodological niceties in efficacy studies, may turn out to be worse than useless for the investigation of the actual treatment of mental illness in the field. It is worth mulling over what the results of an efficacy or effectiveness study might be if half the patients with a particular disorder were randomly assigned and were compared with half the patients not randomly assigned. Appropriately assigning individuals to the right treatment, the right drug, and the right sequence of techniques, along with individuals' choosing a therapist and a treatment they believe in, may be crucial to getting better.

The Ideal Study

The *CR* study, then, is to be taken seriously—not only for its results and its credible source, but for its method. It is large-scale; it samples treatment as it is actually delivered in the field; it samples without obvious bias those who seek out treatment; it measures multiple outcomes including specific improvement and more global gains such as growth, insight, productivity, mood, enjoyment of life, and interpersonal relations; it is statistically stringent and finds clinically meaningful results. Furthermore, it is highly cost-effective.

Its major advantage over the efficacy method for studying the effectiveness of psychotherapy and medications is that it captures how and to whom treatment is actually delivered and toward what end. At the very least, the *CR* study and its underlying survey method provides a powerful addition to what we know about the effectiveness of psychotherapy and a pioneering way of finding out more.

The study is not without flaws, the chief one being the limited meaning of its answer to the question "Can psychotherapy help?" This question has three possible kinds of answers. The first is that psychotherapy does better than something else, such as talking to friends, going to church, or doing nothing at all. Because it lacks comparison groups, the *CR* study only answers this question indirectly. The second possible answer is that psychotherapy returns people to normality or more liberally to within, say, two standard deviations of the average. The *CR* study, lacking an untroubled group and lacking measures of how people were before they became troubled, does not answer this question. The third answer is "Do people have fewer symptoms and a better life after therapy than they did before?" This is the question that the *CR* study answers with a clear "yes."

The *CR* study can be improved upon, allowing it to speak to all three senses of "psychotherapy works." These improvements would combine several of the best features of efficacy studies with the realism of the survey method. First, the survey could be done prospectively: A large sample of those who seek treatment could be given an assessment battery before and after treatment, while still preserving progress-contingent treatment duration, self-correction, multiple problems, and self-selection of treatment. Second, the assessment battery could include well-normed questionnaires as well as detailed, behavioral information in addition to more global improvement information, thus increasing its sensitivity and allowing it to answer the return-to-normal question. Third, blind diagnostic workups could be included, adding multiple perspectives to self-report.

At any rate, *Consumer Reports* has provided empirical validation of the effectiveness of psychotherapy. Prospective and diagnostically sophisticated surveys, combined with the well-normed and detailed assessment used in efficacy studies, would bolster this pioneering study. They would be expensive, but, in my opinion, very much worth doing.

REFERENCES

Consumer Reports. (1994). Annual questionnaire.

Consumer Reports. (1995, November). Mental health: Does therapy help? pp. 734–739.

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

Howard, K., Orlinsky, D., & Lueger, R. (1994). Clinically relevant outcome research in individual psychotherapy. *British Journal of Psychiatry, 165,* 4–8.

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

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies. *Archives of General Psychiatry, 32,* 995–1008.

Muñoz, R., Hollon, S., McGrath, E., Rehm, L., & VandenBos, G. (1994). On the AHCPR guidelines: Further considerations for practitioners. *American Psychologist, 49,* 42–61.

Seligman, M. (1991). *Learned optimism.* New York: Knopf.

Seligman, M. (1994). *What you can change & what you can't.* New York: Knopf.

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

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