Allegiance in Psychotherapy Outcome Research: Separating Association From Bias

Yan Leykin, University of California, San Francisco
Robert J. DeRubeis, University of Pennsylvania

Concern about the contamination of psychotherapy outcome studies by "allegiance bias"—distortion of findings because of investigators' preferences—has led to the proposal that findings to date should not be used to make inferences about the relative efficacies of psychotherapies. It has also been proposed that results from all such studies should be adjusted to cancel the presumed distorting effects of allegiances. We argue that although much effort has been devoted towards establishing the existence of statistical associations between allegiances and outcomes, the causal implication—that investigators' allegiances influence results—has gone virtually untested. We present a new vocabulary with the aim of sharpening the allegiance discourse, and we propose that research strategies markedly different from the ones used to date are needed to address some of the more serious limitations of allegiance bias research.

Key words: **allegiance, allegiance bias, randomized controlled trial, treatment outcome.** ***[Clin Psychol Sci Prac 16: 54–65, 2009]***

BRIEF HISTORY OF ALLEGIANCE

The term *therapeutic allegiance of the experimenter* was first used by Luborsky, Singer, and Luborsky (1975) in their classic systematic review of findings from outcome studies of psychotherapy. They asked peers of the authors of the studies included in their review to estimate the authors' allegiances to the treatments they studied. Although Luborsky et al. did not report any quantitative allegiance-related findings, they made two observations that relied on the estimates obtained from authors' peers. First, they noted that the strongest allegiance estimates were assigned to authors of studies that compared behavior therapy or client-centered therapy to other therapies. Second, they noted that comparisons of behavior therapy to other therapies had all been conducted by researchers with allegiance to behavior therapy, and that likewise, studies that compared client-centered therapy to other therapies had all been conducted by authors with allegiance to client-centered therapy. They expressed the concern that "despite care in design, the therapeutic allegiance of the experimenter might in some way influence the results" (Luborsky et al., 1975, p. 1003). They tempered this caution with the speculation that allegiance would probably cause a problem only in non-double-blind and poorly conducted studies.

Subsequent investigations of the effects of allegiance have used meta-analytic methods, and they have included pharmacotherapy conditions, a broader sample of psychotherapies, and higher-quality studies than were available to Luborsky et al. (1975; e.g., Gaffan, Tsaousis, & Kemp-Wheeler, 1995; Luborsky et al., 1999, 2002; Robinson, Berman, & Neimeyer, 1990). Findings from several of these meta-analyses have indicated statistical relationships between the allegiances of the investigators and the outcomes they report. Moreover, when meta-analysts have controlled for investigator allegiance statistically, they have observed a reduction in the magnitude of between-treatment effect size estimates, leading to the speculation that in treatment outcome research, allegiance distorts or

Address correspondence to Yan Leykin, University of California, San Francisco, 3333 California St., Suite 465, San Francisco, CA 94143-0848. E-mail: Yan.Leykin@ucsf.edu.

© 2009 American Psychological Association. Published by Wiley Periodicals, Inc., on behalf of the American Psychological Association.
All rights reserved. For permissions, please email: journalsrights@oxon.blackwellpublishing.com

biases the comparisons of one treatment with another in favor of the treatment to which the investigator holds an allegiance.

Allegiance has since been described as a "powerful source of bias in comparative outcome research" that not only contaminates the results but also "underscores the futility of comparative outcome designs" (Shoham & Rohrbaugh, 1999). While this view of the effects of allegiance may be more extreme than the norm, many authors, including Lambert (1999) and Jacobson (1999) have warned that allegiance can result in substantial bias. Though emphasizing the importance of efforts to prevent or reduce its effects, Luborsky et al. (1999) have specifically recommended that results from clinical trials should be statistically corrected for allegiance effects, in the service of reducing or eliminating the contamination presumed to result from the allegiance of investigators.

IMPLICATIONS OF ALLEGIANCE BIAS

Insofar as allegiance bias exists in an outcome research domain, the validity of randomized controlled trials (RCTs)—the gold standard in the field for evaluating treatments—would be seriously undermined in that domain. The results of published trials, especially, but not only, those conducted before the implementation of more rigorous standards of research (e.g., CONSORT statement; Begg et al., 1996) would be highly suspect. Indeed, some researchers, citing the findings related to the possible influence of allegiance, have suggested that the RCT as a psychotherapy research method is fundamentally flawed and therefore has little or no value (Luborsky, 2001; Messer & Wampold, 2002).

Thus, if findings from RCTs are to be trusted, the methods used in treatment outcome research may need to undergo fundamental changes, with the aim of limiting or eliminating the biasing effects of allegiance. If, despite current methodological practices, the primary investigator's allegiance still affects the behavior of therapists, evaluators, the research team, and the patients, more careful controls may need to be established, methodologically and statistically. At the extreme of controlling for allegiances, clinical scientists could outsource their studies to independent, dispassionate, neutral researchers, if such could be found. These researchers would then recruit teams of experts to administer each of the treatments being compared.

Given the prospect of discounting or ignoring decades of outcome research, as well as the specter of implementing a new set of untested rules and research guidelines, it is worthwhile to examine critically the existing research on allegiance, and to evaluate the validity of conclusions derived from it.

ALLEGIANCE, ALLEGIANCE ASSOCIATION, AND ALLEGIANCE BIAS

Because it is common in the literature for concepts such as "allegiance," "association of allegiance and outcome," and "allegiance bias" to be used interchangeably, it is important to distinguish these concepts, as we attempt to do in the following.

Allegiance

"Allegiance," in the context of treatment outcome research, is a belief in the superiority of a treatment. It usually also entails a belief in the superior validity of the theory of change that is associated with the treatment. It is unsurprising that allegiance is commonly observed in an investigator who pursues a research question. Smith, Glass, and Miller (1980) found that only in 12% of studies were the allegiances of authors either "balanced" or "unclear." Although allegiances to treatments or to theories may be common or even ubiquitous, it does not follow that the presence of an allegiance will necessarily impact research findings.

Allegiance Association

An "allegiance association" is a statistical relation between the direction and magnitude of treatment outcome findings, and the measured allegiance of the investigators. Allegiance associations have been reported in several meta-analytic studies covering a variety of disorders and their treatments, including depression (Robinson et al., 1990), various anxiety disorders (Berman, Miller, & Massman, 1985), nocturnal enuresis (Houts, Berman, & Abramson, 1994), and schizophrenia (Paley & Shapiro, 2002). Some meta-analyses have encompassed a variety of treatments, and a variety of disorders (Luborsky et al., 1999, 2002; Shirk & Russell, 1992; Smith et al., 1980). The finding of an association between allegiance and outcome is typically instantiated as a significant correlation between the effect sizes of the difference between the outcomes of two treatments and the allegiance scores

assigned to those who conducted and published the research on the treatments (e.g., Luborsky et al., 1999, 2002; Robinson et al., 1990). The correlation between allegiance and outcome superiority has been reported to be as high as 0.85 (Luborsky et al., 1999). Others (e.g., Smith et al., 1980) have examined the differences in pre–post effect sizes of treatments to which researchers had allegiances, noting that these were larger than those for which researchers had no apparent allegiances. Additionally, Luborsky et al. (2002) observed that after controlling for allegiance scores, the effect sizes of the difference between treatments decreased, buttressing the authors' claim of an association between allegiance and outcome.

Not every reviewer of psychotherapy treatment literatures has reported a strong or significant association between allegiance and outcomes. Gaffan, Tsaousis, and Kemp-Wheeler (1995) performed a reanalysis of a previously published meta-analysis (Dobson & Franche, 1989) as well as a separate meta-analysis of more recently published trials of cognitive therapy for depression. They reported little to no association of allegiance to outcome, as well as lower allegiance scores in more recent trials, as compared to the older trials.

Allegiance Bias

The presence of a statistical relation between allegiance and outcome does not necessarily indicate the presence of an "allegiance bias." Unlike "allegiance association," "allegiance bias" connotes a particular causal direction in the interpretation of a statistical association between allegiance and outcome; specifically, that investigators' allegiances (not, for example, the inherent superiority of one treatment over another) are responsible for the advantages found for their preferred treatments. To date, no investigations of allegiance effects have been conducted using experimental or even quasi-experimental methods, so that none of the findings of association between allegiance and outcome can help distinguish between two plausible interpretations of the data: (a) that allegiance has biased outcomes; or (b) that treatments that are superior (relative to the treatments against which they are compared, in the population in which the comparison is carried out) have cultivated the allegiances of investigators. Weiss and Weisz (1995) have made the point that the reverse direction of causality is not only plausible, but perhaps even more likely. They suggest that an allegiance can develop when an investigator observes evidence, either in the clinic or in the literature, of the superior performance of a particular treatment approach for a given problem or disorder.

Berman et al. (1985) reported positive associations between allegiances and outcomes in their meta-analysis, even when the allegiance assessments were based on information from a publication that was published before the conduct of the trial from which data were obtained for the meta-analysis. Referencing this finding, Luborsky et al. (1999) stated that "allegiances affect outcome, not the other way around" (p. 97). Although the Berman et al. (1985) finding shows that the specific outcomes of the constituent studies could not have caused the observed allegiances, there was nothing in their method that provides support for Luborsky and colleagues' (1999) assertion in preference to the following alternative—that qualities of the treatments themselves are responsible both for the investigators' allegiances and for the relative success of these treatments in subsequent trials. Nonetheless, several writers (e.g., Berman et al., 1985; Robinson et al., 1990) have recommended a statistical correction for a problem that, we argue, has by no means been established. Luborsky et al. (1999) count among their strongest recommendations that meta-analysts of outcome literatures should "correct the results for the impact of the researcher's therapy allegiance to each of the treatments compared" (p. 103). In agreement with Klein (1999), we are concerned that such recommendations do not take into account the possibility that there are superior treatments that could affect both the allegiances of outcome researchers and the findings they report.

The observations that investigators have allegiances or that allegiances are associated with outcomes should, therefore, have little effect on one's confidence in treatment outcome literatures, absent evidence that outcome findings result from investigators' preferences, as opposed to the possibility that the inherent qualities of the treatments result both in investigators' beliefs about their effectiveness and the effectiveness of the treatments themselves. Investigations of allegiances and outcomes that are designed to distinguish the plausible causal connections would guide interpreters of outcome literatures appropriately. Findings that favor the "allegiance influences results" hypothesis should lead to greater skepticism of RCTs, whereas findings that fit better with the "good treatments lead to allegiance and good outcomes" hypothesis should increase confidence

in findings from RCTs, as well as in the outcome research enterprise.

MEASURING ALLEGIANCE

Allegiance researchers have relied on four methods to measure the presence, as well as the level, of allegiance: analysis of the publication for the presence of attributes that may hint at allegiance (reprint analysis); similar examinations of previous publications authored by the investigator to establish the presence of preexisting allegiance (before the commencement of the study); interviews with colleagues of the investigators of interest; and interviews with the investigators themselves. With a few exceptions, allegiance investigations have used a single method, reprint analysis. As explained below, most of these methods are unlikely to yield a valid measure of investigator allegiance.

Gaffan et al. (1995) described their procedure for documenting the presence—and strengths—of allegiances. Although most other allegiance investigators (e.g., Houts et al., 1994; Robinson et al., 1990) give much less detail about their procedure, a reasonable assumption is that the Gaffen et al. (1995) procedure is a codification of similar procedures used by other groups. Gaffan and colleagues' indicators are (a) having a reference to a previously published research showing one therapy to be superior to some other treatment; (b) a rationale for a superiority of one therapy over others (i.e., a hypothesis stating that one treatment will outperform the others); (c) a description of the therapy being 10 or more lines long; (d) authorship of the therapy; and (e) having that therapy as the sole active treatment because it is regarded as superior to other available treatments.

Although these criteria may identify the presence and strength of investigators' allegiance, they may, additionally or alternatively, reflect the bona fide superiority of the treatment in question. To the extent that some treatments are superior to others, presenting previous research indicating the superiority of that treatment, or a compelling theoretical rationale to that effect, is indicative of familiarity with the relevant research literature. Longer description of a treatment may be necessary if the treatment is manualized, or relatively novel, or if the intended audience is deemed by the authors (or the papers' reviewers) to be relatively unfamiliar with the treatment. It is also not clear why having a therapy as a sole treatment that is regarded as superior to others is a measure of allegiance. A single-treatment study of an a priori superior treatment is unlikely to be conducted to establish its superiority to a placebo, but rather to address a specific question (e.g., via a dismantling design), or to extend the treatment to a different population, or to test a new vehicle of treatment delivery. The only criterion from the Gaffan et al. (1995) list that appears, on the face of it, to point to a worrisome form of allegiance is the authorship of the treatment under investigation. Indeed, because manuscripts tend to be written after the completion of the study, the findings presumably influence the manner in which data are presented. Reprint analysis is therefore likely to capture an indication of study results, rather than preexisting allegiances. Examination of investigators' previous publication record (see Berman et al., 1985) as an adjunct to the common reprint analysis does ensure that allegiance indicators existed before commencement of the investigation of interest. However, this procedure does not address the other uncertainties that remain after the Gaffan et al. (1995) procedures are used.

Characterizations of investigators' allegiances based on Luborsky and colleagues' (1999, 2002) procedure, which includes the use of interviews with investigators' colleagues, is subject to many of the same sources of uncertainty as are the more common, and less labor-intensive, reprint reviews. As Luborsky et al. (1999) mention, an especially influential published work may bias colleagues' perceptions of the allegiance of the investigator in question. Additionally, personal interests and allegiances of colleagues may likewise bias their assessments. For example, an outspoken critic of Treatment A may judge an investigator who is in fact indifferent between treatments A and B to be an advocate of Treatment A.

Interviews with the investigators themselves constitute perhaps the most straightforward method for ascertaining their allegiances. According to the findings of Luborsky et al. (1999), researchers tend to rate themselves as rather "middle-of-the-road" in terms of their allegiances, with similar ratings towards all treatments under consideration. This may be due in part to the desires of the investigators to uphold the scientific value of empiricism, as opposed to partiality, as well as to the difference between having an affinity towards a particular treatment, and considering it superior to others. For example, an investigator might be disposed towards cognitive therapy for depression because it

appeals to his or her understanding of the nature of the disorder, but he or she may also believe that psychodynamic therapy is an equally effective treatment for depression.

The validity of allegiance ascertainment methods is also called into question by the low correlations among them reported in Luborsky et al. (1999), though it is still, to our knowledge, the only published attempt of this kind. In this study, reprint analysis, by far the most commonly used method, correlated especially poorly with self-reported allegiances ($r = 0.10$).

LIMITATIONS IN ALLEGIANCE RESEARCH TO DATE

Allegiances May Reflect Nature

As we have noted, demonstrations that outcome investigators possess allegiances, and that these allegiances are associated with the direction and magnitude of effects reported in the outcome literature, are not the same as demonstrations of bias. Indeed, if one were to assess the beliefs of allegiance investigators in regard to the allegiance bias hypothesis (allegiance towards allegiance, if you will) and one were to find a correspondence between their beliefs and the findings of their investigations, this would not establish the fact that their allegiances have influenced, or produced, the results they have reported, and therefore the conclusions they have reached. This is because allegiance to a treatment or to a hypothesis, even the allegiance bias hypothesis, may align with nature. One may possess an allegiance to a treatment that is in fact excellent, or better than other treatments, either by chance or because one has astutely identified the outstanding qualities of the treatment. Likewise, it is possible that those who favor allegiance bias interpretations of the findings of outcome research are correct. However, just as one must be cautious about the findings generated by outcome researchers who hold allegiances to treatments, skepticism about the claims of those who warn about allegiance bias is likewise warranted.

Gaffan et al. (1995) and Smith et al. (1980) provide evidence that investigators who are judged to have allegiances to a treatment tend to report larger pre–post effect sizes of that treatment, relative to investigators with lower (or no) allegiance to that treatment. One explanation for this tendency is that positive or negative allegiances cause treatments to yield unrepresentative results. The other possibility, given that those investigators with allegiance to a treatment are likely to be experts in that treatment, is that their studies tend to include more effective implementations of the treatment (Gaffan et al., 1995; Klein, 1999; Shirk & Russell, 1992). If one adjusts the results of a treatment trial to account for the allegiance of an investigator, one risks underestimating the effects that a treatment will produce when it is delivered well.

Allegiance and Comparisons

Luborsky et al. (1999) stated that one of the most common reasons that allegiance bias occurs is that investigators often select a less effective treatment against which to compare their preferred treatment. They include this as a justification for their proposal that results of studies must be adjusted for the investigators' allegiances. However, it does not follow that the results of such comparisons reflect a bias that needs to be controlled. Suppose that an investigator has conducted a study comparing a preferred treatment (A) with a nonpreferred and less effective treatment (B), and reported an effect size estimate, in favor of A, that is quite high (e.g., > 1.0, in d-type terms). Luborsky and colleagues' (1999, 2002) recommendation would be to adjust the effect size estimate downward, to address the fact that a comparison of A to a treatment that is stronger than B (let us call it C) would probably result in a smaller effect size estimate than was obtained in the A versus B comparison. However, there is no need for such an adjustment, as long as the inferences about the effects of A are made in relation to those obtained with B. Comparisons with treatments that are known or presumed to be weak are valid comparisons that can reveal real and important differences between the two treatments tested (Klein, 1999). Such comparisons can by no means be assumed to reflect bias on the part of the investigators.

The assumption that good results indicate bias (and only bias) is also evident in the inferences derived from correlational meta-analyses of allegiance and outcome. For example, although Robinson et al. (1990) included in their dataset randomized trials that compared two plausible treatments head-to-head, they also included comparisons of plausible treatments versus various control conditions (e.g., placebo, treatment-as-usual) or treatments that are widely believed to be ineffective for the disorder under study (e.g., applied relaxation for

obsessive-compulsive disorder). They argued that to the extent that investigators' preferences for treatments (allegiances) predicted (correlated with) the sizes of the advantages of treatments, relative to comparison conditions, one could infer that allegiance bias was at play. They therefore inferred that the significant correlation they obtained ($r = 0.58$) reflected substantial allegiance bias. An at least equally plausible source of such a statistical relation, however, is that comparisons between two treatments that differ greatly in their effects (e.g., a good treatment versus a weak one, or a treatment versus a not-treatment) will yield larger differences than comparisons between two good treatments. Therefore, the magnitude of such correlations does not tell us about the existence of, or influence of, "bias."

Grouping of Diagnoses and Treatments

Luborsky et al. (1999) write,

> It has been perplexing over the years that research comparisons of the same treatments have reported different findings. For example, Gallagher and Thompson (1982) reported a difference between dynamic and behavioral therapies in favor of behavioral; in comparison, Cross, Sheehan, and Khan (1982) reported a difference in favor of dynamic. For the comparisons of cognitive versus behavioral treatments, Butler, Fennell, Robson, and Gelder (1991) found a difference in favor of cognitive therapy while Wilson, Goldin, and Charbonneau-Powis (1983) reported a difference in favor of behavioral therapy. And so it goes in many other studies. Maybe there are hidden factors operating that explain these differences. (p. 95)

The perplexity lessens when one considers that Gallagher and Thompson (1982) studied depression in elderly adults, whereas Cross et al. (1982) studied "disturbed, but not in an extreme sense" (p. 104) clients, wherein the nature of the disturbances was not reported. As for the second comparison, Butler et al. (1991) studied patients with generalized anxiety disorder, whereas the Wilson et al. (1983) study dealt with depressed patients. (We also note that Cross et al. (1982) reported and concluded that there were no differences between behavioral and dynamic treatments in their study. Likewise, Wilson et al. (1983) reported no difference between cognitive therapy and behavioral therapy.) One cannot reasonably draw conclusions regarding allegiance bias from these sets of studies, even if one were to agree with Luborsky and colleagues' summaries of the authors' findings. Rather than resort to "hidden factors" to explain whatever differences might obtain between the pairs of studies, we suggest a more straightforward possibility: that different therapies are better suited to (and thus more effective with) different problems (Siev & Chambless, 2007). To cite an analogy from medicine, if Tylenol is found to outperform Allegra in the treatment of headache, and the reverse is true for seasonal allergies, one would not conclude that hidden factors, or allegiances, are responsible for this pattern in the data.

Similarly, pooling very different kinds of treatments together to form a category can lead to similar problems (Beutler, 2002). For example, in Luborsky and colleagues' (2002) analysis, the categories were as follows: cognitive, behavioral, cognitive-behavioral, and dynamic therapies, and pharmacotherapy. This categorization put together antidepressant medications and antianxiety medications. Likewise, they grouped together two behavioral treatments for depression, applied relaxation and behavior activation, that are quite different procedurally and in their presumed mechanisms of action. It must be considered that a researcher who believes that behavioral activation will outperform applied relaxation for depression may be revealing his or her familiarity with empirical data (see Chambless & Ollendick, 2001), rather than an unjustified allegiance that has influenced the data in his or her outcome study.

Sources of allegiance bias. Thus far we have argued that the empirical data brought to bear on the question of allegiance bias have generally not been of the appropriate kind. This is not to say that allegiances could not have effects that need to be understood, controlled for, or prevented. We now propose some possible effects of allegiance which should be considered and examined, wherever possible.

(1) Allegiance bias is more likely to the extent that a researcher is poorly trained, or deficient in his or her practice of ethics. Researchers who select their primary outcome measure post hoc, or who are not

careful to ensure the random assignment of patients to groups, allow for their partiality to affect the estimates of treatment effects obtained in their studies. One hopes, and we would guess, that egregious examples of this kind are rare if not nonexistent in modern studies of treatments, but subtler versions could indeed affect reported results in unfortunate ways. For example, an investigator who has not planned, and committed to, the details of his or her hypothesis tests in advance could engage in self-deception and judge that a transformation of the data, or a particular way of handling missing data, is the correct way, when it happens that these methods confer advantages to the investigator's favored treatment. Recognition that such effects can occur is reflected in the requirement of many scholarly journals that investigators must have declared their data analytic method before the initiation of a treatment study, if they wish to have the study reviewed for possible publication (De Angelis et al., 2004).

(2) Another possible source of bias is the enthusiasm of the researcher for a particular treatment, leading him or her to engage in behavior that favors the preferred treatment, such as developing and implementing a better training and supervision program for the preferred treatment.

(3) An investigator, after analyzing data from a treatment study, might not pursue vigorously the publication of the findings if they are weakly positive or null (see Luborsky et al., 1999). (Failing to pursue publication of a finding that favored the investigator's nonpreferred treatment would fall into category 1 above.) It is difficult to know how many "file drawer" studies of this kind have been conducted in the past, but it has become increasingly difficult if not impossible for the modern outcome researcher to keep such findings out of the literature. Modern psychotherapy trials generally require substantial funding, so that even if the investigator does not make his or her colleagues aware of the conduct of the study, the funding agency is likely to insist that the results are made public, even if the findings appear to be inconclusive (see Thase, 1999; Klein, 1999).

(4) Finally, bias may occur via "honest differences" in expertise (Hollon, 1999). For instance, the investigator may pursue the goal of conducting a balanced study, but given his or her greater knowledge of the preferred therapy, the performance of the nonpreferred therapy might suffer. This type of allegiance bias can occur despite the best intentions of the investigator, manifesting itself in a variety of ways, including offhand comments to staff or participants, thereby creating an atmosphere that could shift the balance in the direction of the preferred treatment (Jacobson, 1999). This type of allegiance bias is probably widespread and not easily detectable. Although the impact of this type of allegiance may alter the results to a lesser extent than could the other forms of bias, findings that are influenced by it may still affect the discourse about treatment effects.

The divisions noted above are somewhat arbitrary in that there could be a continuum of bias ranging from unethical behavior to innocent behavior that nonetheless could reduce the validity of conclusions made from an outcome literature. The effects of all but the unethical forms of bias-producing behavior can be eliminated or reduced substantially if the field is wary of the possibility of bias, and cognizant of its most likely sources. Several authors (Hollon, 1999; Jacobson, 1999; Luborsky et al., 1999; Thase, 1999) have discussed the importance of collaborations between investigators who possess complementary areas of expertise, and correspondingly opposite allegiances. This approach to research has been termed "adversarial collaboration" by Mellers, Hertwig, and Kahneman (2001). Findings from studies that have exemplified this approach have been published from comparisons of psychotherapeutic versus pharmacotherapeutic treatments (e.g., Barlow, Gorman, Shear, & Woods, 2000; DeRubeis et al., 2005; Foa et al., 2005; Heimberg et al., 1998), but comparisons of different approaches to psychotherapy (e.g., cognitive therapy versus dynamic therapy) have been less common (e.g., Svartberg, Stiles, & Seltzer, 2004).

Several multisite trials (e.g., Barlow et al., 2000; Foa et al., 2005; Heimberg et al., 1998) have utilized adversarial collaboration, with one of the sites well-known for its psychotherapy, and another similarly well-known for its pharmacotherapeutic treatment. The multisite approach is especially advantageous for determining the possible effects of allegiances on outcome within that study. With variables such as inclusion/exclusion criteria and treatment implementation being the same between the sites,

any site differences vis-à-vis treatment superiority can be attributed to allegiance effects. Of the above studies, only one (Foa et al., 2005) reported a significant site-by-treatment interaction for one of the outcome variables. As a single data point, no one such study can be conclusive in relation to the entire treatment outcome enterprise. Nonetheless, this is the type of research method that, along with the approach outlined below, will provide the data that will ultimately lead the field towards either supporting or disproving the allegiance bias hypothesis in the interpretation of outcome findings.

BETTER TESTS FOR ALLEGIANCE BIAS ARE NEEDED

The common method of testing for allegiance bias is the "mixing bowl" design, in which all studies, regardless of comparison groups or diagnoses, are analyzed together. As we have argued, this type of analysis does not allow the analyst to rule out plausible rival hypotheses to the claim that allegiances, per se, have influenced the outcome results. Suppose that a researcher with allegiance to dynamic therapy publishes findings from a small study that compared the effects of dynamic therapy versus an attention-placebo condition in an agoraphobic sample. Further suppose that a cognitive therapy advocate publishes results from a large-sample comparison of cognitive therapy versus dynamic therapy for depression. Results from the first study reveal a strong effect of dynamic therapy, relative to the control condition, and results from the second study favor cognitive therapy, relative to dynamic therapy. In both cases, the results are consistent with the allegiance bias hypothesis, but all that has been demonstrated is an allegiance association. No matter how many studies such as these an allegiance meta-analysis would encompass, it should do little if anything to persuade a skeptical scientist that allegiance bias does in fact account for the results.

It is therefore necessary to develop alternative approaches to test for allegiance biases that avoid or minimize the problems that have plagued the analyses performed to date, and that would allow one to make cautious causal claims regarding the presence and magnitude of the effect of researchers' allegiances on the outcomes of clinical trials. It must be acknowledged, however, that a true experiment, the only research design that would resolve this question definitively, is impracticable at best, and perhaps not possible. A true experiment of the allegiance bias hypothesis would require the random assignment of researchers to treatments, as well as to allegiances to those treatments. A feasible if imperfect alternative would be to approximate the goals of a proper experiment, using quasi-experimental methods. As an example of one quasi-experimental method, pairs or sets of comparative outcome studies would be identified that differ primarily in the allegiances of the investigators. The studies would need to be matched, at least, on the patient population (e.g., adults with panic disorder and agoraphobia) and on at least two of the treatments that were compared (e.g., cognitive therapy versus behavior therapy). The two studies in a given pair would differ in the allegiances of the respective investigators and research teams. Ideally, the pairs of studies would also be matched on other characteristics that plausibly affect outcomes or, more crucially, differences in outcomes between the two treatments. These would include symptom variables, such as severity and chronicity; demographic variables, such as age and socioeconomic status; and treatment variables, such as frequency of sessions, the use of manuals, and therapists' experience and expertise. Multisite trials using the adversarial collaboration approach, as described above (e.g., Barlow et al., 2000; Foa et al., 2005; Heimberg et al., 1998), are, by design, examples of this approach, but the analysis of allegiance bias effects can be conducted, with much greater confidence than is warranted with the correlational designs prevalent in allegiance association findings to date, with pairs of studies that mimic the most important features of multisite adversarial comparative studies. That is, estimates of the effects of allegiance are most likely to reflect the presence and magnitude of bias in adversarial collaborations, and insofar as pairs of studies contain the features of such multisite studies, findings in regard to allegiance bias will also be informative, and relatively free of the limitations of allegiance bias research to date.

Although each of the measures of allegiance have identifiable limitations, thorough interviews with investigators would appear to be the most straightforward and defensible approach, as it avoids the overreaching assumptions and conclusions of other methods while maintaining face validity. The interview itself need not be longer than several questions, which might assess for personal treatment preferences for a particular diagnosis, and/or beliefs about their relative efficacy at the time the

Table 1. Possible outcomes from pairs of studies and their interpretations

	Outcomes of the pair of studies, where investigator allegiance is opposite		Interpretation of result	
	Investigator holds allegiance to A	Investigator holds allegiance to B	Allegiance bias evident?	Relative benefits of treatments
1	A = B	A = B	No	A = B
2	A > B	A > B	No	A > B
3	A > B	A < B	Yes	A = B?
4	A > B	A ≥ B	Yes	A > B
5	A < B	A > B	No	Uncertain
6	A ≥ B	A >> B	No	A > B?

Note: ">>" indicates a difference between treatments greater than that in the other study of the same pair.

study was taking place. An aggregate score would be required for trials conducted by multiple investigators who differ in the strength or direction of their allegiances.

Six results are possible from any pair of studies, or from multisite adversarial collaborations (see Table 1). Cases 1 and 2 represent pairs of studies that provide no support for an allegiance bias, and therefore should promote confidence, all else equal, in the findings from each of the studies. In the first case, the effects of the two treatments are not distinguished by studies from either camp, leading to a tentative inference that the two treatments are similar in the magnitude of their effects. In the second case, both studies find the same treatment to be superior to the other, supporting the hypothesis that one treatment is superior to the other.

The patterns represented by the third and fourth cases would suggest the operation of allegiance bias. In the third case, the studies find opposing results consistent with the respective allegiances of the investigators. This is the pattern that would be expected if allegiance biases are large and treatment differences are small; it would support the inference that allegiance biases were operating and it would add little to an understanding of the relative merits of the two treatments. In the fourth case, the same treatment is superior in both studies, but the magnitude of the effect is larger in the study conducted by investigators with allegiance to the superior treatment. Such a pattern would lend support both to the treatment superiority hypothesis and an allegiance bias hypothesis. The fifth and sixth cases, in which allegiances would appear to have had the effect opposite to expectation, are likely to be rare. Indeed, Luborsky et al. (1999) remarked that they found no studies that pointed to such a negative effect of allegiance.

Undoubtedly, as pairs of studies that admit to this approach accrue, analysts of these literatures will face complex situations, such as pairs of studies that are not perfectly matched, or locating more than two studies that can be compared. While it is our intention to provide only general guidelines, rather than specific answers, some possible solutions for such problems are noted below. In those cases where studies within a pair cannot be perfectly matched, the choice may be to either exclude that particular pair, or weight that data point such that it exerts less influence on the overall pool or study pairs. In situations where there are more than two possible matches, an investigator may either select the best possible match, or, in the case of a multisite trial, include all possible pairs, perhaps as part of a nested analysis. We concur with Luborsky et al. (1999, 2002) that the most trustworthy investigations of allegiance bias will be conducted by teams of analysts that are balanced in regard to the allegiances of the members of the team. In particular, such a team would need to include clinical scientists who believe that investigator allegiance is a serious problem in the existing data, as well as those who are highly skeptical of this claim. Again, the model of adversarial collaboration would be appropriate, as it is in the conduct of clinical trials of competing treatments.

CONCLUSIONS AND RECOMMENDATIONS

To date, there are no conclusive data to either support or reject the allegiance bias hypothesis, in the outcome literature writ large, or in subsets of it. Likewise, there

have been few if any studies that have used an analytic strategy that would permit investigators to draw causal conclusions from any demonstrated association between the outcome of clinical trials and allegiances of the investigators.

It is possible to envision and conduct better tests of the effects of allegiance on outcome. The approach we have proposed addresses a number of methodological limitations of analyses of the associations between allegiance and outcome. Although we propose quasi-experiments, rather than true experiments, studies carried out in a manner consistent with our proposal can nonetheless provide information that appropriately informs judgments about the causal direction of any observed allegiance association. Future investigators are encouraged either to build upon this approach, or to create new ones that also address the limitations of the existing allegiance research.

Although the influence of allegiance on treatment outcome findings has yet to be demonstrated, several recommendations that have been suggested by allegiance researchers should be implemented, both to strengthen research designs and to reduce the magnitude or likelihood of even the unwitting influences of investigators' allegiances. One hopes that competent, ethical researchers, even if they possess an allegiance to one of the treatments they are studying, will attempt to conduct a fair test of the effects of their favored treatment against a control or comparison condition. However, despite good intentions, the expertise and experience of an investigator may not be equal for both treatments. Luborsky et al. (1999) suggested including in the research team persons with allegiance to treatments other than that preferred by the primary investigator, thus balancing the allegiances. Commentators have agreed with and expanded upon this suggestion (Hollon, 1999; Jacobson, 1999; Lambert, 1999; Thase, 1999). Its implementation would ideally take into account not only the allegiances of the members of the research team, but also their corresponding levels of expertise in the approaches to which they evidence allegiance. Assembling a research team in this manner may not only lead to a more balanced investigation, but it may also communicate to the research team the primary investigator's investment in the less-preferred treatment, perhaps lessening the impact of the allegiance-prone "overall ambience of the research site" (Jacobson, 1999, p. 118).

The role of allegiance bias in outcome findings is a controversial and important topic that should neither be dismissed nor embraced without careful attention to the means of assessing it. Writers who have warned about the possible biasing effects of allegiance done the field a service by increasing the awareness of this potential problem. The associations between allegiances and outcomes in various literatures have been estimated, but the direction of the causal relation between allegiances and outcomes has not been given sufficient attention. We have pointed to and recommended the kinds of research methods that can be used both to minimize any biasing effects of allegiance, and to test these effects in a way that goes beyond the finding of a statistical association between allegiance and outcome.

REFERENCES

Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2000). Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *Journal of the American Medical Association*, *283*, 2529–2536.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, *276*, 637–639.

Berman, J. S., Miller, R. C., & Massman, P. J. (1985). Cognitive therapy versus systematic desensitization: Is one treatment superior? *Psychological Bulletin*, *97*, 451–461.

Beutler, L. E. (2002). The dodo bird is extinct. *Clinical Psychology: Science and Practice*, *9*, 30–34.

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

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Cross, D. G., Sheehan, P. W., & Khan, J. A. (1982). Short- and long-term follow-up of clients receiving insight-oriented therapy and behavior therapy. *Journal of Consulting and Clinical Psychology*, *50*, 103–112.

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2004). Clinical trial registration: A statement from the International Committee of Medical

Journal Editors. *Journal of the American Medical Association*, *292*, 1363–1364.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., et al. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, *62*, 409–416.

Dobson, K., & Franche, R.-L. (1989). A conceptual and empirical review of the depressive realism hypothesis. *Canadian Journal of Behavioral Science*, *21*, 419–433.

Foa, E. B., Liebowitz, M. R., Kozak, M. J., Davies, S., Campeas, R., Franklin, M. E., et al. (2005). Randomized, placebo-controlled trial of exposure and ritual prevention, clomipramine, and their combination in the treatment of obsessive-compulsive disorder. *American Journal of Psychiatry*, *162*, 151–161.

Gaffan, E. A., Tsaousis, I., & Kemp-Wheeler, S. M. (1995). Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *63*, 966–980.

Gallagher, D. E., & Thompson, L. W. (1982). Treatment of major depressive disorder in older adult outpatients with brief psychotherapies. *Psychotherapy: Theory, Research and Practice*, *19*, 482–490.

Heimberg, R. G., Liebowitz, M. R., Hope, D. A., Schneier, F. R., Holt, C. S., Welkowitz, L. A., et al. (1998). Cognitive behavioral group therapy vs phenelzine therapy for social phobia: 12-week outcome. *Archives of General Psychiatry*, *55*, 1133–1141.

Hollon, S. D. (1999). Allegiance effects in treatment research: A commentary. *Clinical Psychology: Science and Practice*, *6*, 107–112.

Houts, A. C., Berman, J. S., & Abramson, H. (1994). Effectiveness of psychological and pharmacological treatments for nocturnal enuresis. *Journal of Consulting and Clinical Psychology*, *62*, 737–745.

Jacobson, N. S. (1999). The role of the allegiance effect in psychotherapy research: Controlling and accounting for it. *Clinical Psychology: Science and Practice*, *6*, 116–119.

Klein, D. (1999). Dealing with the effects of therapy allegiances. *Clinical Psychology: Science and Practice*, *6*, 124–126.

Lambert, M. (1999). Are differential treatment effects inflated by researcher therapy allegiance? Could Clever Hans count? *Clinical Psychology: Science and Practice*, *6*, 127–130.

Luborsky, L. (2001). The meaning of "empirically supported treatment" research for psychodynamic and other long-term therapies. *Psychoanalytic Dialogues*, *11*, 583–604.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, *6*, 95–106.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., et al. (2002). The dodo bird verdict is alive and well—mostly. *Clinical Psychology: Science and Practice*, *9*, 2–12.

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies. Is it true that "everyone has won and all must have prizes"? *Archives of General Psychiatry*, *32*, 995–1008.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects: An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

Messer, S. B., & Wampold, B. E. (2002). Let's face facts: Common factors are more potent than specific therapy ingredients. *Clinical Psychology: Science and Practice*, *9*, 21–25.

Paley, G., & Shapiro, D. A. (2002). Lessons from psychotherapy research for psychological interventions for people with schizophrenia. *Psychology and Psychotherapy*, *75*, 5–17.

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

Shirk, S. R., & Russell, R. L. (1992). A reevaluation of estimates of child therapy effectiveness. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 703–709; discussion 710.

Shoham, V., & Rohrbaugh, M. (1999). Beyond allegiance to comparative outcome studies. *Clinical Psychology: Science and Practice*, *6*, 120–123.

Siev, J., & Chambless, D. L. (2007). Specificity of treatment effects: Cognitive therapy and relaxation for generalized anxiety and panic disorders. *Journal of Consulting and Clinical Psychology*, *75*, 513–522.

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

Svartberg, M., Stiles, T. C., & Seltzer, M. H. (2004). Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for Cluster C personality disorders. *American Journal of Psychiatry*, *161*, 810–817.

Thase, M. E. (1999). What is the investigator allegiance effect and what should we do about it? *Clinical Psychology: Science and Practice*, *6*, 113–115.

Weiss, B., & Weisz, J. R. (1995). Letter to the editor. *Journal of the American Academy of Child and Adolescent Psychiatry*, *34*, 971–972.

Wilson, P. H., Goldin, J. C., & Charbonneau-Powis, M. (1983). Comparative efficacy of behavioral and cognitive treatments of depression. *Cognitive Therapy & Research*, 7, 111–124.

Received October 13, 2007; revised July 31, 2008; accepted August 1, 2008.