COMMENTARY

Allegiance Effects in Assessment: Unresolved Questions, Potential Explanations, and Constructive Remedies

Scott O. Lilienfeld and Meredith K. Jones,
Emory University

The provocative results of Blair, Marcus, and Boccaccini (2008) suggest that the allegiance effect, previously suggested in psychotherapy outcome studies, may apply to studies of actuarial risk assessment. Despite this finding, the mechanisms of the effect, particularly in assessment research, are unknown and warrant further investigation. We discuss the file drawer effect, selective reporting, and "data massaging" as three potential explanations for allegiance effects in the assessment domain. Furthermore, we offer four suggestions for minimizing allegiance effects and their impact: routinely coding for allegiance in meta-analytic studies, operationalizing allegiance in multiple ways, encouraging collaborations among authors with differing allegiances, and creating study registries to track all dependent variables measured in studies.

Key words: **allegiance effect, meta-analysis, risk assessment, SORAG, Static-99, VRAG.** ***[Clin Psychol Sci Prac 15: 361–365, 2008]***

Address correspondence to Scott O. Lilienfeld, Department of Psychology, Emory University, Room 207, 532 Kilgo Circle, Atlanta, GA 30322. E-mail: slilien@emory.edu.

Like most important research, the provocative and skillfully executed article by Blair, Marcus, and Boccaccini (2008) on allegiance effects in actuarial risk assessment raises many questions. We intend this observation as a compliment, as their article not only offers the first reasonably clear evidence for "allegiance effects" in psychological assessment, but also points to a host of largely neglected questions that merit attention.

REPLICABILITY AND GENERALIZABILITY

As the late David Lykken, the PhD mentor of the first author, wisely noted, a general principle in science is that the more interesting a finding, the less likely it is to replicate (see Lykken, 1968, 1991, for discussions). After all, intriguing findings are intriguing precisely because they often run counter to conventional wisdom, previous findings, or both. Thus, as a next step, it will be crucial to ascertain that the Blair et al. (2008) findings are replicable and generalizable to assessment measures other than actuarial risk devices. Moreover, assuming that these findings are replicable, it will be essential to delineate their boundary conditions. Are allegiance effects larger for measures that allow examiners more "leeway" and subjective judgment in probing for responses, coding responses, interpreting responses, or all three? If so, as Blair et al. (2008) observe, the effects they reported for actuarial risk measures may underestimate the magnitude of the effects "for measures requiring a greater level of clinical judgment in administration and scoring" (p. 347). If not, this negative finding would itself be informative, as it would suggest that allegiance effects operate not at the stages of administration, scoring, and interpretation, but at the stages of data analysis, reporting, or both.

ALLEGIANCE: WHAT IS IN A NAME?

As in the psychotherapy outcome literature (e.g., Luborsky et al., 1999), the term "allegiance effect" may be something of a misnomer, as it implies a bias, intentional or unintentional, on the part of test developers to accrue supportive evidence for their pet measures. As a consequence, this phrase may imply a deeper understanding of the causal processes involved than is warranted. In many respects, the finding of an allegiance effect is more of a phenomenon in need of explanation than an explanation per se. Given our ignorance of the reasons for allegiance effects, the more neutral term "investigator effect" may be warranted, although we will not lobby strongly for this preference (despite our allegiance to it!).

Allegiance effects may, indeed, reflect either explicit or implicit biases on the part of investigators, but they may also reflect expertise effects arising from superior knowledge of an assessment instrument, as Blair et al. (2008) acknowledge. For example, researchers who possess more intimate familiarity with a measure that they have constructed may obtain higher quality information, code participants' responses more accurately, or design studies with a better eye to ruling out subtle confounding variables. Being more motivated to generate positive results for their preferred measure, they may also institute more rigorous procedures to minimize rater drift over time, resulting in higher inter-rater reliability for measures that require ratings by observers. Blair et al. wisely examined the possibility of superior methodological quality among studies by allegiant versus nonallegiant investigators by incorporating several methodological moderators (e.g., retrospective versus prospective design) into their meta-analysis, but as they point out, this possibility is difficult to exclude with certainty. Absence of evidence is not evidence of absence.

ALLEGIANCE EFFECTS: THREE POTENTIAL EXPLANATIONS

Assuming that the Blair et al. (2008) results are replicable and generalizable to other assessment measures, which remains to be seen, their results raise a number of intriguing explanations for allegiance effects (in addition to differences among studies in methodological quality, mentioned earlier) that merit further investigation. We discuss three such possibilities here (see also Meehl, 1990; Rosenthal, 1994, for discussion of potential mechanisms underpinning expectancy effects in the educational literature).

File Drawer Effects

Blair et al. (2008) note that the "file drawer effect," the tendency for negative findings to remain unpublished (Rosenthal, 1979; see also Cooper, DeNeve, & Charlton, 1997; Staines & Cleland, 2007), may be one explanation for assessment allegiance effects. Indeed, in general, published investigations tend to yield effect sizes that are approximately one-third larger than those of unpublished investigations (Conn, Valentine, Cooper, & Rantz, 2003). Moreover, in the medical literature, effect size differences between published and unpublished studies persist even after controlling statistically for differences between the two classes of studies in methodological quality (Stern & Simes, 1997). File drawer effects can themselves result from at least two separable biases: a failure on the part of authors to submit negative reports ("submission bias") or a failure on the part of journals to accept them ("publication bias"; Cooper et al., 1997; Meehl, 1990). Presumably, only the former tendency would generate allegiance effects.

Selective Reporting

We suspect that another important contributor to allegiance effects is selective reporting of dependent measures *within* studies. Regrettably, this bias has received considerably less attention than the file drawer effect (which operates across studies), perhaps because it is more difficult to detect. As assessment researchers know, construct validation studies frequently incorporate a large array of dependent measures, and decisions regarding which measure(s) to report in published articles are often matters of subjective judgment or even stylistic preference. Allegiance effects toward a measure may shape such decisions, even if unwittingly.

In the treatment outcome literature, this effect is sometimes referred to as "outcome reporting bias" (Chan & Altman, 2005). Its magnitude is difficult to estimate, but provisional evidence suggests that it may be substantial. For example, Chan and Altman (2005) found that about 20% of outcome measures in medical trials went unreported, and that unreported outcomes were about twice as likely to be statistically nonsignificant as significant. Interestingly, among authors who insisted

that they had reported all outcome measures in their published reports, 32% nevertheless mentioned at least one outcome measure in their *Method* section that went unreported in their *Results* section.

The magnitude of selective reporting of dependent measures in validational studies of psychological assessment measures is unknown, and should become a major priority for future researchers. In our own literature reviews of the validity of the Rorschach inkblot test (Lilienfeld, Wood, & Garb, 2000; Wood, Lilienfeld, Garb, & Nezworski, 2000), we identified at least one example in which an author had apparently neglected to report multiple negative findings in a published report, despite presenting them in his dissertation. Of course, such selective reporting may, in certain cases, be a reasoned decision guided by methodological considerations (e.g., the findings for the unreported measures may be of questionable methodological quality) or by journal page constraints. Nevertheless, the absence of reported information on null findings can distort markedly the verdicts of meta-analytic reviews.

Data "Massaging"

As Lykken (1991) observed,

> The processes of planning, conducting, and analyzing any psychological experiment are complicated, frequently demanding decisions that are so weakly informed by any ancillary theory or established practice as to seem essentially arbitrary. As the investigator makes his or her way through this underbrush, there is the ever-beckoning lure of the desired or expected outcome that tends to influence the choices made at each step. (p. 8)

In the Byzantine world of data analysis, researchers must often make a host of subjective decisions at multiple stops along the long and winding road to publication: whether to exclude outliers and if so, which criteria to use for doing so; whether to exclude participants on the basis of apparent invalid responding, and if so, which scales and which cutoffs to use; whether to transform skewed data and if so, how; whether to report results for the full sample or to instead report results for subsamples (e.g., males and females, White people and African Americans) when differences in correlational patterns are of ambiguous magnitude and importance; whether and how to impute missing data; whether to report interactions of questionable meaning or magnitude or to pool them with error terms; and so on. Any or all of these seemingly subtle decisions, most or all of which may be made in good faith, can stem from confirmation bias (Gilovich, 1991; Nickerson, 1998) and inadvertently paint an overly rosy picture of an assessment measure's validity.

Of course, researchers may also engage in "data dredging," that is, post hoc exploratory analyses of data conducted in an effort to locate statistically significant findings (Joseph & Baldwin, 2000). Data dredging is sometimes defensible as long as investigators acknowledge in published reports that the analyses were exploratory. But when investigators misleadingly present exploratory analyses as confirmatory (i.e., as hypothesized a priori), a spurious inflation of effect sizes can result.

Selective data checking (Lykken, 1991; Meehl, 1990) can further contribute to allegiance effects. When the validation findings for a measure confirm our predictions, we are unlikely to go back to our raw data or program statements to spot possible errors. In contrast, when findings run counter to our predictions, we are more likely to do so (e.g., "did we perhaps forget to recode those two reversed items on the scale?"), resulting in a bias toward positive results. Investigators who are motivated to see their preferred measure cast in a favorable light may be especially prone to such post hoc double-checking.

CONSTRUCTIVE REMEDIES

The findings of Blair et al. (2008) point to a number of constructive remedies for potential allegiance effects in the assessment literature. We would maintain that such remedies, four of which we present here, are advisable even if these allegiance effects prove to be limited to only a subset of assessment instruments.

Routinely Coding for Investigator Allegiance in Meta-Analyses

One implication of the Blair et al. (2008) findings is straightforward and should be relatively noncontroversial: Researchers who conduct meta-analyses on the validity of psychological assessment measures should routinely code for investigator allegiance, and treat such allegiance as a categorical moderator in their analyses. But as in the psychotherapy outcome literature, operationalizing allegiance can be a tricky business.

Operationalizing Investigator Allegiance in Multiple Ways

As a consequence, it may be advisable to operationalize allegiance in multiple ways (see also Luborsky et al., 1999), such as by self-ratings, colleague ratings, and reprint ratings. Moreover, coding the allegiance of not only the first author but also the second author may be advisable given findings in the psychotherapy literature that "investigator divergence"—the difference in allegiance between the first and second authors—correlates negatively with treatment effect sizes (Luborsky et al., 1999; Thase, 1999). Although this finding is open to multiple interpretations, it raises the possibility that research teams with offsetting allegiances to a given assessment measure may be less likely to generate spuriously inflated effect sizes.

Encouraging Collaborations Among Investigators With Differing Allegiances

The investigator divergence findings noted above lead naturally to another suggestion that is relatively noncontroversial but not easily implemented: encouraging collaborations among researchers who differ in their allegiances toward an assessment instrument. As Klein (1999) observed, one of the best antidotes against allegiance effects in the psychotherapy literature is the formation of multidisciplinary research teams consisting of members with varying theoretical orientations. In the assessment domain, finding investigators with markedly divergent allegiances toward the Rorschach inkblot test and other projective techniques, for example, is hardly difficult, although getting these investigators to agree to collaborate (or even to talk to each other) may be more of a challenge. Before embarking on such collaborations, researchers from differing assessment camps should ideally agree up front on "joint ground rules" for the design of studies and on the results that would either corroborate or disconfirm claims of validity for the measures of interest (see Hyman & Honorton, 1986, for a model of a joint communiqué authored by investigators with opposing allegiances in a markedly different area of psychological research).

Establishing Study Registries

As many authors have observed, one of the best antidotes against file drawer effects and selective reporting is the establishment of a publicly accessible registry of all Institutional Review Board (IRB)-approved studies (Cooper et al., 1997; Simes, 1986). Such a registry would facilitate the unearthing of the large "fugitive" or "gray" literature (Joseph & Baldwin, 2000) of unpublished or difficult-to-access findings in the assessment domain, which could militate against allegiance effects stemming from publication biases. Indeed, there is at least some evidence that studies retrieved from publicly accessible registries of medical data yield smaller effect sizes than those of studies accessed from traditional medical databases, like *Medline* (Cooper et al., 1997).

Nevertheless, this suggestion may only be effective to the extent to which authors report all dependent variables in their study proposals. To the extent that authors add measures following the approval of studies or create new operationalizations of constructs in a post hoc fashion in data analyses (e.g., by aggregating two or more measures that are highly correlated into a composite index), the establishment of study registries may only minimize, but not eliminate, the problems posed by unreported findings.

CONCLUDING THOUGHTS

Scientific methods are essential, albeit imperfect, safeguards against a multitude of sources of bias (O'Donohue, Lilienfeld, & Fowler, 2007; Tavris & Aronson, 2007). By identifying a heretofore largely unappreciated source of potential error in the psychological assessment literature, Blair et al. (2008) have done the field a valuable service. If their findings turn out to have broader import for the assessment field at large, they will pose an important challenge to consumers of the literature. At the same time, their findings may point to significant biasing influences that, when adequately controlled, may clarify unresolved debates regarding the validity of controversial psychological measures.

REFERENCES

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice*, *15*, 346–360.

Chan, A., & Altman, D. G. (2005). Identifying outcome reporting bias in randomized trials on PubMed: Review of publications and survey of authors. *British Medical Journal*, *330*, 753.

Conn, V. S., Valentine, J. C., Cooper, H., & Rantz, M. J. (2003). Grey literature in meta-analyses. *Nursing Research*, *52*, 256–261.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447–452.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.

Hyman, R., & Honorton, C. (1986). A joint communique: The psi ganzfeld controversy. *Journal of Parapsychology*, *50*, 351–364.

Joseph, J., & Baldwin, S. (2000). Four editorial proposals to improve social sciences research and publication. *International Journal of Risk and Safety in Medicine*, *13*, 117–127.

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

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

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 & Practice*, *6*, 95–106.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology. Volume 1: Matters of public interest* (pp. 2–39). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

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

O'Donohue, W. T., Lilienfeld, S. O., & Fowler, K. A. (2007). Science is an essential safeguard against human error. In S. O. Lilienfeld & W. T. O'Donohue (Eds.), *The great ideas of clinical science: 17 principles that every mental health professional should understand* (pp. 3–27). New York: Routledge.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, *3*, 176–179.

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology*, *4*, 1529–1541.

Staines, G. L., & Cleland, C. M. (2007). Bias in meta-analytic estimates of the absolute efficacy of psychotherapy. *Review of General Psychology*, *11*, 329–347.

Stern, J. M., & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal*, *315*, 640–645.

Tavris, C., & Aronson, E. (2007). *Mistakes were made (but not by me): Why we justify foolish beliefs, bad decisions, and hurtful acts*. Orlando, FL: Harcourt.

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

Wood, J. M., Lilienfeld, S. O., Garb, H. N., & Nezworski, T. M. (2000). The Rorschach tests in clinical diagnosis: A critical review, with a backward look at Garfield (1947). *Journal of Clinical Psychology*, *56*, 395–430.

Received April 29, 2008; accepted April 29, 2008.