

## Models and meanings: Therapist effects and the stories we tell

STEPHEN SOLDZ

*Boston Graduate School of Psychoanalysis*

*(Received 10 June 2004; revised 5 October 2002; accepted 21 December 2004)*

Two studies, same data set, opposite conclusions. Therapist effects are either vital or virtually non-existent. Previous research either supports the importance of therapist effects or provides little or no evidence that these effects exist. What kind of science is this?

The question the articles in this special section address is an important one for psychotherapy research: Is there major variation in outcomes between therapists? Since the pioneering work of Crits-Christoph et al. (Crits-Christoph, Baranackie, Kurcias, & Beck, 1991; Crits-Christoph & Mintz, 1991), psychotherapy researchers have been aware that therapist effects can be a potential confound in studies of treatment effectiveness. Leaving these effects out of analyses can result in incorrect parameter estimation and statistical tests and the drawing of incorrect inferences. More practically, the existence of substantial therapist differences in effectiveness could lead to the examination of the characteristics and practices of the more effective therapists as contrasted with those of the least effective therapists; the resultant knowledge could help improve psychotherapy practice.

Both articles in this special section, by Elkin, Falconnier, Martinorich, and Mahoney, and Kim, Wampold and Bolt, examine the existence of therapist effects in the Treatment of Depression Collaborative Research Program (TDCRP) data set. Both use multilevel models, the relatively new, and generally preferred, analytic technique for analyzing clustered hierarchical data. Yet they come to contrary conclusions.

In examining these two studies, the first thing that strikes me is their quality. Both are excellent examples of psychotherapy research.<sup>1</sup> They are closely reasoned, methodologically sophisticated, and carefully executed. They both contain details on their modeling strategy, including the presence in

each of an appendix with more Greek letters than many psychotherapy researchers are used to.

In comparing these studies, it is important not to overestimate the substantive difference in findings. Elkin *et al.* found no variables exhibiting significant therapist effects and found between 0% and 4% of the variance in outcomes explained by therapist differences. Wampold found a significant therapist effect for only one variable, the Beck Depression Inventory (BDI), and estimated the magnitude of therapist effects at between 0% and 11% of the variance in the analyses using the same sample as Elkin *et al.* Thus, those strongly attached to null hypothesis significance testing (Cohen, 1994) need only confront a differing result for one variable. Those willing to contemplate nonsignificant effect size estimates confront a somewhat larger discrepancy.

In searching for the reasons for the divergent findings, one can look at potential differences in the samples, in measures, or in statistical analysis. Whereas Wampold analyzed two samples, his intent-to-treat sample contains the same 119 patients who Elkin *et al.* included. Thus, sample differences are not responsible.

As for measures of outcome, the studies overlap in using three measures (Hamilton Rating Scale for Depression, Global Assessment Scale, BDI). Elkin *et al.* used two more measures of social and work functioning. Wampold also used Hopkins's Symptom Checklist-90. Thus, there is substantial overlap in measures. Fortunately, the only measure to actually exhibit a statistically significant therapist effect in Wampold, the BDI, was included in both studies. In the Wampold model with both random therapist intercepts and slopes, therapist effects explained 8.3% of the variance in BDI outcome (Table 5;  $p < .001$ ) in the intent-to-treat sample, whereas in Elkin *et al.* therapist effects explain 2.3%

---

Correspondence: Stephen Soldz, Boston Graduate School of Psychoanalysis, 1581 Beacon Street, Brookline, MA 02446. E-mail: ssoldz@bgsp.edu

of the variance in BDI. Thus, differences in measures are not the cause of the discrepant findings.

We are thus left with statistical analyses as the likely culprit. Here we find major differences. Both authors adopted a hierarchical linear models (Raudenbush & Bryk, 2002)—also known as multi-level (Heck & Thomas, 2000; Hox, 2002; Kreft & de Leeuw, 1998; Snijders & Bosker, 1999) or mixed-effects models (Pinheiro & Bates, 2000)—approach to analysis.<sup>2</sup> These models correct for the nonindependence (in a statistical sense) of nested observations. Thus, patients seen by the same therapist may have more similar outcomes than would two patients chosen at random. Further, if one wishes, one can include outcome measures at multiple time points during the course of treatment, creating a growth-curve model (Fitzmaurice, Laird, & Ware, 2004; Rogosa, 1995; Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003; Willett, 1997).

Wampold used a two-level multilevel model, with random therapist effects nested within treatments (a fixed effect) and patients nested within therapists as a random effect. The dependent variable was the outcome variable assessed at termination. Elkin et al. added a third level to Wampold's model structure, nesting the multiple repeated measures of each outcome variable within the patient level. As with Wampold, the patients were nested within therapists, who, in turn, were nested within treatments as a fixed effect.

Whereas Wampold assessed patient outcome using the termination value (or last available measurement for study dropouts) of each dependent variable (with the pretest score for that variable entered as a predictor), Elkin et al. assessed patient outcome by the rate of change (on a log-linear scale) in their repeated measures. Thus, the two studies used different definitions of outcome; as a result, their estimates of the magnitude of therapist effects are not directly comparable.

A further difference between models is that Wampold included the baseline measure as a covariate in his models, as has become traditional in psychotherapy research. Elkin et al., in contrast, adopted an anchored approach, in which each patient's growth curve is forced to go through that patient's baseline assessment, thus eliminating one source of variability in the models.

### Modeling

Barring error,<sup>3</sup> the difference in results between the two studies resides in the choice of model. We thus are confronted with a set of issues that other quantitative disciplines face, but that psychotherapy research, and the disciplines from which it primarily

draws (psychology, social work, psychiatry, psychoanalysis), have been largely, and blissfully, ignorant. These issues fall under the general statistical approach to data analysis referred to as modeling (Krzanowski, 1998). Like Molière's Monsieur Jourdain, who spoke prose all his life without knowing it, psychotherapy researchers have long been engaged in modeling largely without awareness. So does this lack of awareness matter? I argue that it matters insofar as it has kept us isolated from a large body of literature on modeling and its complexities from which we can learn (e.g., Greene, 2002; Harrell, 2001; Kennedy, 2003). Further, I argue that use of data analytic techniques such as multilevel models requires us to understand a range of issues of which most of us psychotherapy researchers<sup>4</sup> have remained blissfully unaware.

Among the topics and issues central to the modeling literature but largely absent from our graduate research methods and statistics curricula are the nature of estimators, the variety of estimation algorithms, model selection criteria, the testing of model assumptions, and the consequences of the violation of assumptions. Thus, how many of us understand the differences between ordinary least squares, maximum likelihood, and robust variance estimators, much less understand Bayesian and empirical Bayes approaches? Do we consider the effects that differing algorithms can have on model parameter estimation (de Leeuw & Kreft, 1994, 1995)? How many of our studies using linear growth curves exhibit an awareness of the potential consequences of nonlinear change<sup>5</sup> or of other potential misspecifications of models? Do our studies exhibit an awareness of the role of shrinkage in the estimation of an individual patient's growth curve? Have we pondered the consequences of using multilevel models with data sets quite a bit smaller than those for which asymptotic results can be assumed to hold and smaller than those recommended by some statisticians (Kreft & de Leeuw, 1998)? How many of our studies report an examination of residuals or other tests for misspecification? Were those tests conducted? Are those currently embracing multilevel models aware of alternatives, such as generalized estimating equations (Hardin & Hilbe, 2003), that may be at least as appropriate for certain studies?

These issues may seem abstruse, but the adoption of multilevel models forces us to confront them. These, and other contemporary advanced statistical techniques, are capable of leading to quite wrong answers if not used carefully and thoughtfully. Because these analyses are complex and their output is somewhat foreign, it can also be harder to tell whether the analysis has gone awry. The point here is not to discourage use of these techniques, because

they constitute a decided increase in our analytic capabilities and allow us to address questions that we previously could hardly imagine. However, their use requires care, because we are not in Kansas anymore. In many cases, as with the studies in this Special Section, appropriate use of them will require that a member of the research team be a statistician or a researcher who has expended considerable effort to become familiar with contemporary statistical thought and practice. As we move toward the multilevel era, the days of research as a cottage industry conducted by a single researcher, with perhaps the assistance of a graduate student who took an extra statistics course, may be numbered.

So what is the model that these techniques estimate? It consists of an idealized view of a presumed mechanism underlying the production of the data at hand. The model is idealized in that only a subset (in psychotherapy research, a small subset) of potential variables is included, and the nature of the relationship between dependent and independent variables (e.g., linearity or direct vs. interaction effects) is only a first approximation. All models, like all scientific laws, have an implicit *ceteris paribus* assumption, “other things being equal,” which, of course, never actually holds in the real world, even the world of physics (Cartwright, 1983), much less that of psychotherapy.

Models, being idealizations, are not true representations of reality. Just as different artists, or even the same artist at different times, can paint the same scene differently, so can different models encompass different aspects of reality. Thus, there is no such thing as a correct model. Yet that does not mean that everything goes when constructing models. There are various statistical techniques for determining whether one model is superior to another, but choosing among models also involves a pragmatic criterion. Some models are more useful than others (de Leeuw, 1994). In the end, the purpose of our statistics, like that of all our research, is to allow us to tell empirically defensible stories about the world and the way it works (Soldz, 2000).

Relevant to the studies in this Special Section, de Leeuw (1994) emphasized that a prime criterion for good statistical models is that they are robust, not in the sense of robust estimators but that differing analyses using different assumptions come to similar conclusions. Results that vary dramatically in the conclusions we draw, and thus in the stories we tell, are to be less trusted. From this perspective, the conclusion from the differing conclusions of these two studies, given that both were well designed and conducted, is that we cannot be very certain regarding the extent of therapist effects in the TDCRP.

Such effects may exist, but their magnitude is in dispute.

In considering these models as ways of estimating the magnitude of therapist effects, it is important to keep in mind that the TDCRP data set, although one of the larger psychotherapy data sets, is quite small for sustaining the analyses reported here. The techniques used to conduct the analyses reported in these studies are asymptotic in nature, which means that they assume the sample is suitably large, when “large enough” is hard to precisely determine but is often estimated using simulation studies.

In using a small number of therapists and a smaller number of patients per therapist, the analyses of Elkin et al. and Wampold violated the recommendations of many authors on multilevel analysis regarding minimum sample size (Hox, 2002; Kreft, 1996; Maas & Hox, 2002, n.d.). These recommendations suggest, for a two-level model, a minimum of 30 groups; some suggest 30 observations per group. Maas and Hox (2002), pointed out that, with small samples containing only 30 higher level groups (therapists in the studies we are examining), the estimated standard errors for the group (therapist) variance estimates were about 15% too small; with only 10 groups, the standard errors were between 16% and 30% too small, depending on other assumptions. These results suggest that the standard errors obtained by Wampold are likely too small, thus increasing the chance of a Type I error. Thus, the significant results for the BDI should be taken with caution, and the estimates of proportion of variance resulting from therapists are likely less precise than Wampold’s variance estimates and standard errors would lead one to believe.

Wampold acknowledged the low power of his analyses, using this fact as a reason to base his interpretation on the variance estimates rather than traditional significance tests. However, low power increases the standard errors for the estimates of proportion of variance resulting from therapists, leading to greater imprecision in the therapist variance estimates. Elkin et al. also discuss the issue of power in their footnote 3. These authors concluded that there is no known way to estimate power in the models they use. Unfortunately, neither study presents confidence intervals or standard errors for their variance estimates. Thus, we can only guess at their precision, which is presumably not great. Unfortunately, the TDCRP, despite being one of the largest samples of its kind, is actually still not large enough to lead to precise estimates of variability resulting from therapists.

In one of the early studies on therapist effects, Crits-Christoph et al. (1991) conducted a meta-analysis, thus benefiting from combining studies.

Meta-analytic approaches continue to have great potential in this and a number of other areas of psychotherapy research through their ability to increase power by combining multiple studies. Such studies benefit from a relative standardization in the data available from the individual studies they seek to combine. As we move into the multilevel models era, it behooves us to develop minimal standards for how such models should be presented. One factor entering into the creation of such standards is that the information provided should be sufficient for inclusion in a meta-analysis.

### Recommendations for reporting research

In addition to their analysis of the TDCRP data, Elkin et al. made another contribution through their critique of the extant studies on therapist effects. Regardless of the merits of their comments on particular studies, this critique sets an implicit standard for future studies. Among the points raised are that the magnitude of therapist effects often varies considerably depending on what outcome measure is examined, with no clear relationship between measure and variance estimate emerging and with rank-order differences in therapist functioning varying across measures. Elkin et al. also correctly emphasized the importance of examining potential therapist effects for evidence that they may be based on a small number of outlier therapists.

Based on the Elkin et al. analysis, as well as an examination of the current studies examining the TDCRP, some tentative recommendations for future work on therapist effects emerge: (a) All parameter estimates, especially those of therapist variance components, should be accompanied by standard errors, confidence intervals, or some other measure of precision; (b) statistical models should be subjected to examination of model assumptions, such as linearity, and the results of this examination should be noted; (c) analysts should inspect their data for the presence of both therapist and patient outliers and report the results of these inspections; and (d) researchers should give an indication of the degree of consistency of therapist rankings across the various outcome measures and methods of assessing therapist quality used.

Finally, in agreeing with Elkin et al. that the TDCRP is not an ideal data set for examining therapist effects, I concur with their judgment that progress in this area will require large data sets, with many hundreds, if not thousands, of cases, such as those available to managed-care companies, state mental health and substance abuse systems, and those engaged in developing or implementing outcomes monitoring systems (California Children &

Youth Performance Outcome Measurement System, 2001; Center for Substance Abuse Treatment, 1995; Hodges & Wotring, 2004; Lambert, Okiishi, Finch, & Johnson, 1998). With these large databases, it should be possible to estimate the magnitude of therapist effects with at least moderate precision. It remains to be seen whether any of these databases contain enough information regarding therapist characteristics to productively examine those characteristics associated with better therapeutic functioning. Another issue concerns the amount of detailed information regarding patient characteristics, such as the presence of personality pathology, that is necessary for accurately controlling for case-mix differences among therapists. However, for the issue of therapist effects, and for many other issues in psychotherapy research as well, I suspect, the future largely will lie with these large databases. It is hoped that psychotherapy researchers will be involved in the creation of these data systems, so that they can serve our research needs simultaneously with the practical reasons for which they are primarily developed.

As to my opening question, "What kind of science is this?," my answer is this: a science making progress, one becoming refined enough to create conundrums begging for resolution. In other words, a healthy science.

### Notes

- <sup>1</sup> Truth-in-packaging warning: I was a reviewer for one of these studies.
- <sup>2</sup> I personally prefer the term *multilevel model*, which also seems to be the dominant one in the statistical literature. The term *hierarchical linear model* and its acronym *HLM* are less desirable for two reasons. First, HLM is the name of a particular computer program, leading to confusion between software and analytic technique (similar to that in structural equation modeling, which is sometimes referred to as LISREL modeling, after the computer program). Second, these models are being generalized to nonlinear models, including those for nominal and ordinal dependent variables (Fitzmaurice, Laird, & Ware, 2004; Skrondal & Rabe-Hesketh, 2004). We thus have the odd situation in which the HLM (hierarchical *linear* models) program now contains procedures for nonlinear modeling!
- <sup>3</sup> At the risk of offending, I should say that error, although unlikely with such experienced research teams, cannot be completely ruled out. Multilevel models are relatively complicated, and the software to analyze them can be difficult to use. Although both of these studies are by research teams with considerable experience with these models, I am concerned about the tendency for routine use of multilevel models by those less experienced in their use than these authors. Perhaps we are entering the stage at which studies should be required to include computer code and data (when possible), perhaps on the World Wide Web, along with the published study. It would be useful for the authors of these two studies to exchange computer code and even to explore each other's models to locate more precisely the exact reasons for the discrepant findings.
- <sup>4</sup> Although not the authors of these studies, I suspect.

<sup>5</sup> Fortunately, Elkin et al. are well aware of this issue and deal with it successfully.

## References

- California Children & Youth Performance Outcome Measurement System. (2001, January 10). *Children/youth performance outcome monitoring system: Clinical training manual*. Retrieved September 29, 2004, from <http://www.dmh.cahwnet.gov/RPOD/PDF/Child-Training-Manual.pdf>
- Cartwright, N. (1983). *How the laws of physics lie*. Oxford, UK: Oxford University Press.
- Center for Substance Abuse Treatment (1995). *Tip 14: Developing state outcomes monitoring systems for alcohol and other drug abuse treatment*. Washington, DC: Author.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997–1003.
- Crits-Christoph, P., Baranackie, K., Kurcias, J. S., & Beck, A. T. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, 1, 81–91.
- Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting & Clinical Psychology*, 59, 20–26.
- de Leeuw, J. (1994). *Statistics and the sciences*. Unpublished manuscript, Department of Statistics, University of California, Los Angeles. Los Angeles. Retrieved November 19, 2004, from <http://preprints.stat.ucla.edu/download.php?paper=152>
- de Leeuw, J., & Kreft, I. G. G. (1994). *Questioning multilevel models*. Research Triangle Park, NC: National Institute of Statistical Sciences.
- de Leeuw, J., & Kreft, I. G. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics*, 20, 171–189.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience.
- Greene, W. H. (2002). *Econometric analysis* (5th ed). Upper Saddle River, NJ: Prentice Hall.
- Hardin, J. W., & Hilbe, J. M. (2003). *Generalized estimating equations*. Boca Raton, FL: Chapman & Hall/CRC.
- Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag.
- Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Erlbaum.
- Hodges, K., & Wotring, J. (2004). The role of monitoring outcomes in initiating implementation of evidence-based treatments at the state level. *Psychiatric Services*, 55, 396–400.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Kennedy, P. (2003). *A guide to econometrics* (5th ed). Cambridge, MA: MIT Press.
- Kreft, I. (1996, June 25). *Are multilevel techniques necessary? An overview, including simulation studies*. Retrieved March 4, 2005, from <http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html>
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
- Krzanowski, W. J. (1998). *An introduction to statistical modelling*. London: Arnold.
- Lambert, M. J., Okiishi, J. C., Finch, A. E., & Johnson, L. D. (1998). Outcome assessment: From conceptualization to implementation. *Professional Psychology: Research and Practice*, 29, 63–70.
- Maas, C. J. M., & Hox, J. (2002). *Sample sizes for multilevel modeling*. Retrieved December 23, 2003, from <http://www.fss.uu.nl/ms/jh/publist/simnorm1.pdf>
- Maas, C. J. M., & Hox, J. (n.d.). *Robustness of multilevel parameter estimates against small sample sizes*. Retrieved December 23, 2003, from <http://www.fss.uu.nl/ms/jh/papers/p090101.pdf>
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Thousand Oaks, CA: Sage.
- Rogosa, D. (1995). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum.
- Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 92, 726–748.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. London: Oxford University Press.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.
- Soldz, S. (2000, September). *Research as the telling of empirically justified stories*. Paper presented at the meeting of Health and Addictions Research, Boston, MA.
- Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel & K. A. Renninger (Eds.), *Change and development: Issues of theory, method, and application* (pp. 213–243). Mahwah, NJ: Erlbaum.