J Exp Criminol (2007) 3:271–279
DOI 10.1007/s11292-007-9037-x

Unjustified inferences about meta-analysis

Mark W. Lipsey

Published online: 31 July 2007
© Springer Science + Business Media B.V. 2007

Abstract Berk (Statistical Inference and Meta-Analysis, 2007) asserts that the results of inferential statistics make scientific sense only if the data to which they are applied were actually generated through random sampling from a defined real population. Because meta-analysis data are not generated in that manner, he claims that the statistical conclusions of meta-analysis are fictional and suggests that conventional research review procedures be used instead. This rejoinder argues that Berk's position on statistical inference represents a narrow literalism that he fails to justify and that does not reflect the way inferential statistics are used or generally understood in contemporary practice. Consequently, his critique has little significance for meta-analysis or any of the other widespread forms of social science research that apply inferential statistics in similar spirit. Berk's advocacy of conventional literature reviews omits any explanation of how they would avoid the well-documented deficiencies of that approach or be conducted in a manner that offers any advantage over meta-analysis.

Keywords meta-analysis · research synthesis · statistical inference

Unjustified inferences about meta-analysis

Berk (Statistical Inference and Meta-Analysis) asserts that the validity of the inferential statistics applied in meta-analysis depends on how the data were actually generated. In particular, he argues that those statistics will not make scientific sense unless the studies from which effect sizes are computed are randomly sampled from a defined population of studies. Because studies, in fact, are not actually randomly sampled in virtually all meta-analyses, we are told that the conclusions of meta-analysis are a fiction. In light of

M. W. Lipsey (✉)
Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies, Vanderbilt University, 1207 18th Avenue South, Nashville, TN 37212, USA
e-mail: mark.lipsey@vanderbilt.edu

this, Berk suggests that we should not do meta-analysis at all but, instead, conduct conventional literature reviews which "have served science well for a very long time." In other writings he is even more explicit about this view:

> Finally, with respect to meta-analysis, our recommendation is simple: just say no. The suggested alternative is equally simple: read the papers, think about them, and summarize them. (Berk and Freedman 2003).

Berk himself says "no" to meta-analysis. *Evaluation Review*, which he edits, publishes no meta-analyses, despite advertising a focus on "reporting the findings of evaluation studies" and widespread recognition elsewhere that meta-analysis is now the state of the art for summarizing and analyzing such findings. We should take a moment to celebrate the *Journal of Experimental Criminology*, where meta-analyses are not only published but critiques such as Berk's, and rejoinders such as this, are invited so that the issues can be openly engaged. It is through such open engagement within what Donald Campbell (1984) once called the "disputatious community of scholars" that we make progress on the very difficult issues with which we grapple in the applied social sciences.

Meta-analysis is not the only type of research for which statistical inference yields invalid conclusions in Berk's view. He asserts that the significance tests in the individual studies on which a meta-analysis is based are similarly flawed when the treatment and control subjects are not randomly sampled from a large population of subjects, which they rarely are. The implication, one might presume, is that we should not do this type of research either or, if we must do it, at least refrain from applying any statistical tests. By analogy, perhaps we should simply examine the scores for each participant, think about them, and summarize them.

The participants in survey research are often drawn randomly from a defined population, so perhaps Berk would accept that inferential statistics might make scientific sense for those data. But all the individuals sampled in such surveys do not necessarily provide data—indeed, response rates are frequently well below 100%—and that process is not random. In these common cases with attrition, therefore, we do not really have a fully random process generating the data, and, thus, if we apply Berk's standards, we should not conduct statistical significance tests in this research either.

The circumstances that meet Berk's criteria for applying inferential statistics are thus exceedingly rare. If such statistics are applied in any other circumstances (that is to say, pretty much all the research conducted in criminology) Berk would have us believe they have no utility. For meta-analysis, he indicates that description (descriptive statistics?) might be acceptable, at least for those of us who find it hard to say "no." But the basic implication of Berk's argument is that we should not use inferential statistics in all but a few rare instances of criminological research and that their use in meta-analysis is so compromising that we should abandon that enterprise altogether.

What's wrong with this picture?

Berk's position is that of a statistical literalist. In his view the data literally have to be obtained through probability samples drawn from defined real populations for inferential statistics to be applicable or meaningful. Are all the meta-analysts, field

experimenters, and survey researchers who use inferential statistics on data that do not meet this strict criterion as benighted and errant is Berk would have us believe? No, there's quite another way of using statistical inference that characterizes what most of us do and which Berk does not acknowledge at all.

Let's take the simple case of an experiment on treatment for drug offenders conducted with probationers from an urban juvenile court. Suppose we find that the recidivism rate for those receiving the treatment is 0.06 lower than that for those not receiving it. Were the outcomes better for those treated? Within the limits of measurement error, they clearly were for these particular individuals. But we are not especially interested in the idiosyncrasies of these particular individuals. Can we interpret our results to mean that the outcomes for individuals *like* these particular ones would also be expected to show such a difference and, in particular, that they would not show no difference at all or a difference in the opposite direction? Depending on how much natural variability there is among such persons, and how many participated in our experiment, the difference we found might not be representative of what we would find if we did the same experiment over again the same way with a substantially similar group of participants. If not, our finding is too narrowly specific to the particular individuals we studied to be of much general interest.

So, we do a simulation. We suppose that a large population of persons very much like those in our actual study had participated in this experiment, but that the difference in recidivism between those receiving the treatment and those not receiving it in this population was actually zero. We further suppose that we draw samples at random from this population of the same size as we used in our study, and that we do this over and over again. How likely would it be that these samples would generate the data we actually obtained? If this hypothetical scenario indicates that the recidivism difference we found in our study could be readily generated simply by a random draw from a population in which there was no difference, it undermines our confidence that it is replicable. Put another way, we assess the tenuousness of the estimate of the difference we derived from our data against benchmarks defined by a statistical model that tells us how likely it is that such a difference could be spuriously generated by a chance process that assumes replication of the same study with substantially similar persons. We do not need to assume that our data were actually generated by a chance process in order to estimate the probability that such a process would generate those data nor are the results meaningless if they give us some basis for appraising the stability of our finding given the sample size, measures, and distribution of observations on which it was based.

Statistical inference in meta-analysis is done for the same purpose and in the same spirit, but is more complicated because of the multi-level nature of the effect size data. Conducted under the fixed effects model, a test of the significance of a mean effect size asks whether it is likely that such a mean could be generated by the random sampling of the participants for each study from a larger population of very similar such persons in which the effect was zero. In the random effects model we further ask whether it is likely that the observed mean would be generated by a random sample from a population of studies much like the ones in the meta-analysis but with a mean effect size of zero. In neither case does the meta-analyst believe that random sampling was actually done from such populations. Nonetheless, knowing

how readily the observed mean could be generated by such a hypothetical chance process is informative for thinking about the stability of the results obtained.

We conduct these simulations by way of statistical models that integrate probability theory, distributional assumptions, and variance estimates from our empirical samples. The criterion for whether or not the resulting standard errors, tests of statistical significance, and confidence intervals are meaningful is not whether the data were actually generated through the random sampling assumed in the simulation, as Berk insists, but whether they provide an informative basis for assessing the robustness of the findings. While it may be true that too much emphasis has been put on the results of this particular form of assessment in contemporary social science (cf. Cohen 1994), it is not self-evident that it makes no scientific sense. Berk's contrary view is presented as if it were self-evident, and he provides nothing but repetitive assertions to support it.

If we view statistical inference as a simulation of a situation that did not actually occur, of course, it still needs to be a relatively good simulation to be useful. The results of a statistical model that generates inferential statistics on the assumption of independent data points, for instance, will be misleading if the actual data are not independent. Berk makes the point that the effect sizes analyzed in meta-analysis are not independent because of the web of personal contacts and intellectual influences that characterize the researchers who conduct the studies that report those effects. He presents no evidence that the degree of statistical dependency is large enough to produce great misestimation of the standard errors that are central to significance testing, but it is possible that it is. Dependencies of this sort may be widespread in social science data, though difficult to estimate, and arguably deserve more attention. For instance, Berk et. al (2003) analyzed reports of misconduct for inmates in California prisons as independent data points. But we might suppose that inmates in prison participate in a web of relationships such that the misconduct of one, or lack thereof, would influence the misconduct of another in ways that would violate the assumption of independence.

Influences of this sort are difficult to rule out for any data derived from people who are in physical or social proximity to each other. The effect of such dependencies on statistical models that do not account for them is underestimation of standard errors and overestimation of statistical significance. These distortions are recognizable as a form of *design effect* and corrective adjustments are possible if the extent of the dependencies can be estimated. There are also analyses, such as hierarchical linear modeling, that take them into account and are applicable in some instances. This is a topic that deserves more exploration in meta-analysis as well as in Berk's inmate misconduct data and many other areas of research in criminology. The mere suspicion of such dependencies, however, does not automatically render statistical inference so invalid as to be meaningless. Rather, it should motivate efforts to obtain better empirical estimates of the extent of such dependencies, use of statistical models that incorporate them, and, in the nearer term, more conservative interpretations of inferential statistics that do not account for them.

In short, the meta-analyst need concede nothing to Berk's critique other than recognition that improvement may be possible in the statistical models commonly used. His main point is that meta-analysis applies inferential statistics to data that were not actually generated through random sampling and, therefore, the

conclusions based on those statistics are wrong. Taken at face value, this claim has no more applicability to meta-analysis than to most other forms of social science research, few of which involve random samples from defined real populations. Indeed, it has less applicability, because it is typical of meta-analysis to attend mainly to the magnitude of the effect size estimates and place much less emphasis on statistical significance than is common in other areas of research. The origins of meta-analysis owe much to recognition of the limitations of statistical significance for representing the findings of empirical research, and this carries through to present day practice (Schmidt 1992). In any event, if Berk's broad brush tars all research, it paints no picture distinctive to meta-analysis that justifies singling it out for negation.

More to the point, however, is that Berk's argument should not be taken at face value but should, instead, be seen as a reflection of a narrow perspective on statistical inference that lacks a compelling justification (literally, since Berk presents none) and ignores the quite viable alternate view that underlies most current practice. There is nothing in statistical theory that prohibits its use for exploring "what if" scenarios that tell us how readily a chance process of a particular sort could produce our observed results from a hypothetical null result population. The only issues are whether that exercise is useful in helping us appraise the uncertainty with which we should view those results and whether the assumptions made in constructing the statistical scenario match the structure of the data of interest closely enough to produce informative results. On the first point, it may be arguable whether statistical inference on that basis is very informative, but there is nothing about it that renders meta-analysis invalid. On the second point, Berk raises relevant questions about possible interdependencies in meta-analysis data, questions that are also applicable to many other research situations. These suggest that we can improve our statistical models, or may need to make more conservative assumptions to compensate for such design effects, but they do not constitute fatal flaws that justify Berk's "just say no" attitude toward meta-analysis.

And what is the alternative?

Berk proposes to banish meta-analysis from the methodological repertoire for the sin of using statistical inference in ways that his strict statistical literalism does not endorse. It will be in good company—his narrow standard will similarly banish virtually all experiments and quasi-experiments and most survey research. For meta-analysis, however, Berk proposes an alternative: the conventional literature review. We should, he says, simply "read the papers, think about them, and summarize them" (Berk and Freedman 2003). It is tempting to think that Berk offers this suggestion tongue in cheek, with a wink of recognition that it constitutes a reductio ad absurdum argument against his own conclusion. No such wink is evident, however, in the deadpan assertion that such reviews have "served science well for a very long time."

Not mentioned in this glib assertion is the fact that the main impetus behind the development and rapid expansion of meta-analysis is recognition of serious deficiencies in conventional literature reviews, so serious that their conclusions

can easily be completely wrong. This is especially true for reviews of experimental and quasi-experimental studies that investigate whether certain interventions are effective in changing targeted outcomes, the very area in which meta-analysis has been most widely applied. The "vote counting" of statistical significance that is typical of such conventional reviews is demonstrably flawed, especially under the rather typical circumstances where the studies being reviewed have modest sample sizes and correspondingly limited statistical power (Bushman 1994; Hedges and Olkin 1980). Indeed, as Hedges and Olkin (1980) have shown, this technique produces increasingly erroneous conclusions as the number of studies available for review increases, exactly the opposite of what should happen as the body of evidence gets larger.

There is no more dramatic instance of the deficiencies of conventional literature reviews and the corrective influence of meta-analysis than the "nothing works" controversy in criminology over the effectiveness of rehabilitation treatment for offenders. The notorious Lipton et al. (1975) review of 231 studies reported little evidence of positive effects on recidivism, a conclusion echoed by other conventional reviews at the time as well (e.g., Sechrest et al. 1979). What meta-analysis brought to this issue was, first, explicit criteria for defining the studies judged relevant and a thorough search to locate and include all studies meeting those criteria. This reduced the potential for subjective picking and choosing, or collections of convenience, to misrepresent the full body of available evidence. Second, and most important, meta-analysis systematically examined the direction and magnitude of the effects reported across all these studies without being distracted by whether each, individually, was statistically significant within the constraints of its limited sample sizes and circumstances.

The results from meta-analysis were stunning. Every reviewer who applied this more systematic approach to compiling and interpreting the research on offender rehabilitation found that the overall mean effect on recidivism was positive, a complete reversal of the "nothing works" conclusions of the previous generation of conventional literature reviews. Table 1 summarizes these results for the meta-analyses that covered smaller or larger portions of the general research on this topic. In addition, many of the meta-analyses examined the distribution of effects around the mean and showed that some interventions rather consistently produced relatively large effects, further disputing the "nothing works" conclusion. It is worth noting that these meta-analysis results are quite compelling without the tests of statistical significance that Berk disparages. The fact that the mean effect sizes are, in fact, statistically significant, however, bolsters the case. That part of the analysis provides assurance that the statistical estimates these means represent are not highly tenuous, given the number of studies, sample sizes within studies, and within- and between-study variability on which they are based. The story of what meta-analysis contributed to the clarification of the nature of the evidence on the effectiveness of rehabilitation has been told more fully and with more flourish by others (e.g., Cullen 2005; Palmer 1992), but all versions provide an object lesson on how poorly science can be served by the conventional literature reviews Berk advocates in the name of rigor.

Indeed, against this background, we have to ask just what it is that Berk is advocating when he advises us to renounce meta-analysis and do conventional

Table 1 Meta-analyses of the effects of rehabilitation treatment on recidivism

Meta-analysis Report	Age of Offenders	Treatment Setting	Mean Effect Size[a] (N of Studies)	Change in Recidivism[b]
Garrett 1985	Juveniles	Residential	−0.05[c] (19)	−10%
Whitehead and Lab 1989	Juveniles	Community & residential	−0.12[d] (50)	−24%
Andrews et al. 1990	Juveniles & adults	Community & residential	−0.10 (88)	−20%
	Juveniles	Community & residential	−0.10 (70)	−20%
	Adults	Community & residential	−0.11 (18)	−22%
	Juveniles & adults	Community	−0.11 (68)	−22%
	Juveniles & adults	Residential	−0.07 (20)	−14%
Petrosino 1997	Juveniles & adults	Community & residential	−0.10[e](115)	−20%
	Juveniles	Community & residential	−0.12[e] (55)	−24%
	Adults	Community & residential	−0.07[e] (53)	−14%
Cleland et al. 1997	Juveniles & adults	Community & residential	−0.08 (515)	−16%
	Juveniles	Community & residential	−0.08 (288)	−16%
	Adults	Community & residential	−0.07 (227)	−14%
Lipsey and Wilson 1998	Juveniles	Community	−0.13[f](117)	−26%
	Juveniles	Residential	−0.07[f] (83)	−14%
Illescas et al. 2001	Juveniles & adults	Community & residential	−0.17[g] (22)	−34%
	Juveniles	Community & residential	−0.19[g] (13)	−38%
	Adults	Community & residential	−0.10[g] (15)	−20%
Latimer et al. 2003	Juveniles	Community & residential	− 0.09 (156)	−18%

[a] Phi coefficient; unweighted mean when available. A negative sign means less recidivism for the intervention condition. Cohen's *d* effect sizes converted to phi as $phi = d/\sqrt{4 + d^2}$; odds ratios converted to *d* as $d = \log(OR)/2$, then *d* converted to phi (this gives the phi that occurs with a 0.50 control recidivism and the given odds ratio)

[b] Difference between the recidivism rate for the intervention and a control recidivism rate assumed to be 0.50 that corresponds to the given effect size

[c] Subset with random or matched designs and recidivism outcomes

[d] Computed from Table 1 in the original article

[e] Randomized studies only

[f] Unweighted means computed from original data

[g] European studies; subset with controls

reviews instead. To review studies of the effects of interventions, such as those on rehabilitation treatments, he surely doesn't mean using the discredited technique of vote counting the statistical significance of study findings. In Berk's view, the reports of statistical significance for intervention studies are virtually all bogus anyway, so that cannot be what he has in mind. He suggests that "good description" is appropriate and that he would simply summarize what he reads in the studies. What is it he would describe? He might tally up the direction of effects—how many outcomes favor the intervention conditions and how many favor the control conditions, irrespective of statistical significance. But the magnitude of the differences could be quite different in one direction or the other, so he might want to take account of that. If so, how would he describe and assess effect magnitude? He objects to the effect size indices meta-analysts use for this purpose. Would he invent a different index or simply rely on a subjective assessment? Would it matter that some studies had much larger samples and maybe should be given more weight in his summary? Should variation in results associated with different subject samples and method quality be taken into account? How would that information be described

and summarized so that the implications for the conclusions of the review would be revealed? Suppose there are 50 or 100 or more studies to be reviewed. How would Berk keep track of all these particulars? Will his plan to simply read the papers and think about them produce a good representation of what that body of evidence actually adds up to?

If Berk attempts to deal with these matters thoughtfully and systematically, he will simply reinvent meta-analysis. If he reads, thinks, and summarizes, based only on his own impressions and cognitive algebra to produce an old-fashioned narrative review, he will obscure the basis for his conclusions and offer them only as "trust me" assertions that cannot be readily cross-checked or replicated. Is this the answer to his call for greater rigor in research reviews?

What Berk's critique offers on the subject of meta-analysis is a narrow and unconvincing statistical literalism and, apparently, a suggestion that we turn the clock back to the days of unsystematic subjective assessments as our way of summarizing empirical findings. We should just say "no," and say it emphatically.

References

Andrews, D. A., Zinger, I., & Hoge, R. D. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28*(3), 369–404.

Berk, R. (2007). Statistical inference and meta-analysis. *Journal of Experimental Criminology* (in press).

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg, & S. Cohen (Eds.), *Punishment and social control: Essays in honor of Sheldon Messinger* (2nd ed., pp. 235–254). NY: Aldine de Gruyter.

Berk, R. A., Ladd, H., Graziano, H., & Baek, J.-H. (2003). A randomized experiment testing inmate classification systems. *Criminology and Public Policy, 2*(2), 215–242.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–213). NY: Russell Sage Foundation.

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Connor, D. G. Altman, & C., Jackson (Eds.), *Evaluation studies review annual*, (vol. 9, pp. 26–48). Newbury Park, CA: Sage.

Cleland, C. M., Pearson, F. S., Lipton, D. S., & Yee, D. (1997). Does age make a difference? A meta-analytic approach to reductions in criminal offending for juveniles and adults. Presented at Annual Meeting American Society of Criminology, San Diego.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*(12), 997–1003.

Cullen, F. T. (2005). The twelve people who saved rehabilitation: How the science of criminology made a difference. *Criminology, 43*(1), 1–42.

Garrett, C. J. (1985). Effects of residential treatment on adjudicated delinquents: A meta-analysis. *Journal of Research in Crime and Delinquency, 22*(4), 287–308.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*(2), 359–369.

Illescas, S. R., Sanchez-Meca, J. S., & Genovés, V. G. (2001). Treatment of offenders and recidivism: Assessment of the effectiveness of programmes applied in Europe. *Psychology in Spain, 5*(1), 47–62.

Latimer, J., Dowden, C., & Morton-Bourgon, K. E. (2003). Treating youth in conflict with the law: A new meta-analysis. Report RR03YJ-3e. Ottawa: Department of Justice, Canada.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders. In R. Loeber & D. P. Farrington (Eds.) *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 313–345). Thousand Oaks, CA: Sage.

Lipton, D., Martinson, R., & Wilks, J. (1975). *The effectiveness of correctional treatment: A survey of treatment evaluation studies*. NY: Praeger.(1975)

Palmer, T. (1992). *The re-emergence of correctional intervention*. Newbury Park, CA: Sage.

Petrosino, A. (1997). What works? Revisited Again: A Meta-analysis of Randomized Field Experiments in Rehabilitation, Deterrence, and Prevention. Doctoral Dissertation, Rutgers, the State University of New Jersey, Newark.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173–1181.

Sechrest, L. B., White, S. O., & Brown, E. D. (1979). *The rehabilitation of criminal offenders: Problems and prospects*. Washington, DC: National Academy of Sciences.

Whitehead, J. T, & Lab, S. P. (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency, 26*(3), 276–295.

Mark W. Lipsey is the Director of the Center for Evaluation Research and Methodology at the Vanderbilt Institute for Public Policy Studies. His primary research activities involve meta-analysis of research about risk factors for antisocial behavior and the effectiveness of delinquency prevention and intervention programs.