Human Resource Management Review 27 (2017) 193–200

Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

What inferences can and cannot be made on the basis of meta-analysis?

Kevin R. Murphy

Department of Psychology, Colorado State University, 1415, Gen Eagle Ct., Fort Collins, CO 80525, United States

ARTICLE INFO

Article history:
Received 21 March 2015
Revised 12 June 2015
Accepted 13 June 2015

Keywords:
Meta-analysis
Effect size
Variability in effect sizes
Sampling error

ABSTRACT

Meta-analysis involves inferences about the mean, the variability and the substantive reasons for variability in effect sizes. The risks inherent in each of these inferences are reviewed, and a Bayesian approach to using meta-analysis to determine whether effects vary in important ways is suggested.

Meta-analyses have been conducted for over 100 years (Rosenthal & DiMatteo, 2001), but this method for integrating the results of multiple research studies did not become popular in the social and behavioral sciences until the late 1970s (Glass, 1976). Since then, its growth has been astonishing. Between 1994 and 2009, there were over 3000 papers published in psychological journals and over 11,000 in medical journals dealing with meta-analysis (Aguinis, Pierce, Bosco, Dalton, & Dalton, 2011), and this technique for summarizing research literatures continues to grow in popularity and complexity. Meta-analysis is mainly used in integrating the results of multiple studies to develop a quantitative picture of what research on particular interventions or relationships has to say, but in some fields it is moving beyond integrating study results and toward integrating data from multiple large-scale databases (Fox, Lancaster, Laird, & Eikhoff, 2014).

There are several methods for conducting meta-analyses (e.g., Brannick, 2001; Hall & Brannick, 2002; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Pitchforth & Mengersen, 2012; Schmidt, 2007; Schmidt & Hunter, 2003; Schmidt, Oh, & Hayes, 2009; Sutton, Abrams, & Jones, 2000), but they all share a core of common elements. First, a meta-analysis starts by collecting a sample of studies that examine (or claim to examine) comparable research questions. Key study results are then expressed in terms of some effect size indicator that allows researchers to compare and combine results from separate studies that might involve different measures, different research designs and different methods of statistical analysis. Next, some combination of descriptive and inferential statistics is applied to help the analyst make sense of what these studies have to say (and perhaps whether they are all saying similar things). Finally, inferences about what these statistics mean are drawn.

The simple description of meta-analysis in the preceding paragraph belies the many complexities of conducting a meta-analysis. Important judgments must be made about what studies to include in or exclude from a meta-analysis. Meta-analyses that do not adequately sample the relevant literature or that rely on low-quality studies can yield misleading results (Baumeister, DeWall, & Vohs, 2009; Coyne, Thombs, & Hagedoorn, 2010; Stewart & Roth, 2004). Judgments must be made about how to translate the results of studies that use different measures and designs into a common effect size metric. Conducting a meta-analysis can involve numerous decisions in which there is no simple path to identifying the right answer (what to include/exclude, how to represent findings of each study, coding potential moderators, what moderators to consider, what corrections to apply) and differences in the way different

http://dx.doi.org/10.1016/j.hrmr.2015.06.001
1053-4822/

researchers handle these judgment calls can sometimes make a difference in the conclusions meta-analysts will reach (Wanous, Sullivan, & Malinak, 1989).[1]

The studies in a meta-analyses are rarely replications, and they typically employ a range of methods of quantifying key constructs and evaluating their relationships (Aguinis, Pierce, et al., 2011; Rousseau, Manning, & Denyer, 2008), so decisions about which studies should or should not be included and what constructs are actually measured in particular studies are particularly important and complex. For that reason, it is important to make the criteria and rationale used to decide which studies to include in or exclude from a meta-analysis explicit and clear (Aguinis, Pierce, et al., 2011). As will be noted in a later section, the decisions about what studies to include in or leave out from a meta-analysis can have an important bearing on the inferences that can be drawn from that meta-analysis.

Early proponents of meta-analysis were enthusiastic about the role of meta-analysis in uncovering scientific truth. For example, Schmidt (1992) claimed that meta-analysis was necessary and perhaps sufficient to tell us what the data really mean in the social sciences. His analyses suggested that our knowledge regarding some questions (e.g., relationship between perceptual speed and performance of clerical workers) was so precisely established via meta-analysis that it might be irresponsible to continue conducting empirical research on these questions. Subsequent meta-analysts have been a bit more circumspect about what meta-analysis can and cannot accomplish. In order to understand what a meta-analysis can or cannot tell you, it is useful to look carefully at the statistics this method relies on and the samples that are used to make inferences about the effects of interventions or the relationships among variables.

1. Meta-analytic statistics

At its heart, meta-analysis involves some very simple statistics. Suppose, for example, you read 10 studies examining the relationship between perceptions of organizational support and job satisfaction, and you copied down the correlation between support and satisfaction in each study, applying appropriate corrections for the levels reliability and the range restriction in these two variables.[2] You might end up with something like Table 1.

1.1. The mean effect size

Meta-analysis starts (and often pretty much ends) with the simplest of descriptive statistics — i.e., the mean. The simple average of the corrected r values in Table 1 is .379. If you weight the values by sample size average of the corrected r values is .391. As far as many users of meta-analysis are concerned, this mean is all they want or need to know to understand the relationship between two variables or the effects of some intervention (Aguinis, Pierce, et al., 2011). Meta-analyses tend to do a good job of estimating the mean effect size in a group of studies (Oswald & McCloy, 2003; Schmidt, 2007; Schmidt, Oh, & Hayes, 2009) but it is important to consider other aspects of the distribution of effect size estimates in drawing inferences from a meta-analysis.

The interpretation of meta-analytic estimates of effect sizes depends substantially on the extent which the effects of treatments or interventions or the relationships among variables vary across settings, across different methods of defining the treatments or interventions, across different measures, etc. There are probably few research literatures that can be adequately characterized by a single number (the mean effect size), and a statement such as "the correlation between scores on a cognitive ability test and job performance is .50" will almost certainly need qualifications. For example, this particular relationship appears to vary as a function of job complexity (Gutenberg, Arvey, Osburn, & Jenneret, 1983). The sections that follow examine problems in drawing inferences from a meta-analysis when effect sizes vary in substantive ways, but interpreting the mean effect size in a meta-analysis can be a complex undertaking. The meta-analytic mean is a sample statistic that estimates some population parameter. The problem of determining what population and what parameter is not necessarily a simple one.

1.1.1. What is the population?

Suppose the simple meta-analysis described in Table 1 involves ten studies that all use similar measures of the key constructs and all sample a similar mix of members of organizations. It might be reasonable to believe that there is a single effect size that characterizes studies of this sort, making what is referred to as the *fixed-effects* model for meta-analysis appropriate. In a fixed effects meta-analysis, the goal is estimate this common effect size, and once you know the mean effect size in the set of studies you are examining, you know pretty much everything you need to know about the relationship between support and satisfaction in this group of studies. You *might* also be able to infer that in similar studies that were not included in your meta-analysis, or even in similar studies that have not been done yet, this same effect size would still hold, and that in this population of actual and potential studies, the corrected correlation between support and satisfaction is indeed .391. I emphasize *might* here because even if a single effect size does adequately describe what happened in these ten studies, you cannot necessarily be sure that it will also describe the results of other similar studies.

The fixed effects model of meta-analysis is probably useful only in those rare circumstances where the studies in a meta-analysis are all quite similar, and the inferences you are attempting to draw from such a meta-analysis are limited to other studies that

[1] Aguinis, Dalton, et al. (2011), note that there is a very large literature dealing with the judgment calls that need to be made in conducting a meta-analysis. They reviewed the outcomes of "…196 meta-analyses including 5,581 effect-size estimates published in *Academy of Management Journal, Journal of Applied Psychology, Journal of Management, Personnel Psychology,* and *Strategic Management*" (p. 5). On the whole, these seemingly important judgment calls had only small effects on the effect sizes reported in various meta-analyses.

[2] There is a robust literature dealing with statistical and psychometric corrections; for the purposes of this paper, it will be assumed that scientifically-appropriate corrections have been applied, but we will not examine these corrections in any detail.