Neuropsychol Rev (2016) 26:121–128
DOI 10.1007/s11065-016-9319-z

REVIEW

A Guide to Conducting a Meta-Analysis

Mike W.-L. Cheung[1] · Ranjith Vijayakumar[1]

Received: 29 February 2016 / Accepted: 2 May 2016 / Published online: 21 May 2016

Abstract Meta-analysis is widely accepted as the preferred method to synthesize research findings in various disciplines. This paper provides an introduction to when and how to conduct a meta-analysis. Several practical questions, such as advantages of meta-analysis over conventional narrative review and the number of studies required for a meta-analysis, are addressed. Common meta-analytic models are then introduced. An artificial dataset is used to illustrate how a meta-analysis is conducted in several software packages. The paper concludes with some common pitfalls of meta-analysis and their solutions. The primary goal of this paper is to provide a summary background to readers who would like to conduct their first meta-analytic study.

Keywords Literature review · Systematic review · Meta-analysis · Moderator analysis

Most researchers (e.g., Chalmers et al. 2002; Cooper and Hedges 2009; National Research Council 1992; O'Rourke 2007) credit Karl Pearson as one of the earliest users of meta-analytic techniques. Pearson (1904) studied the relationship between mortality and inoculation with a vaccine for enteric fever by combining correlation coefficients across 11 sample studies. The term *meta-analysis* was coined by Gene Glass to represent "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass 1976, p.3). A recent special issue (Shadish 2015) on the origins of modern meta-analysis, published in *Research Synthesis Methods*, provides a comprehensive history of the development of meta-analysis in the social, educational and medical sciences. The widely use of meta-analysis across many disciplines has shown that meta-analysis is a versatile method to synthesize research findings.

There are two main objectives of this paper. The first is to provide an introduction of when to conduct a meta-analysis. Common issues, such as the advantages of meta-analysis over conventional narrative review, and how to decide on the number of studies required for a meta-analysis, are examined. The second objective is to discuss the steps in conducting a meta-analysis. Commonly-used models and software packages, illustrated with an artificial dataset, will then be introduced. It should be noted that this paper is not meant to be comprehensive. Its primary goal is to provide a summary of background knowledge to readers who would like to conduct their first meta-analytic study.

What Is Narrative Review and its Limitations?

One important activity in scientific research is the accumulation of knowledge and research findings (Hunt 1997). Historically, this was done using a narrative review. A researcher would collect information about the studies she finds important or worthwhile regarding a phenomenon, and make inferences about the model by examining the individual studies. She would try to come up with a final judgment about the connection between the variables, by looking at the different studies and evaluating these studies on several criteria. She may also try to analyze the differences between the studies, by looking at particular study sample- or design- features. The

✉ Mike W.-L. Cheung
mikewlcheung@nus.edu.sg

[1] Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570, Singapore

researcher may count the numbers of significant and nonsignificant results using vote counting. If the majority of the results are significant, the researcher may claim that there is some evidence of effect in the studies. If the votes are similar, the researcher may conclude that no conclusive result is observed. Thus, more empirical studies are required.

While narrative review is a step towards creating a more reliable scientific basis for confirming or refuting theories, this approach suffers from a few disadvantages that ultimately render narrative reviews unsatisfactory (e.g., Cooper and Hedges 2009). Firstly, the sheer wealth of information from all disparate studies suggests that the scientists need a justification for focusing on some but not other issues in the articles in the literature. It becomes highly subjective as to which patterns are considered important and which ones irrelevant. Secondly, narrative reviews do not have a systematic tool for combining the results of several studies. Narrative reviews lack a method to merge the findings together to give a more reliable final measure of the result. Their focus on statistical significance rather than the magnitude of effect, is a problem that traditional reviews share with individual studies.

Thirdly, narrative reviews do not adequately correct for sample characteristics, or design features in any explicit fashion. There is no quantification of the effect of study characteristics on the results, except in an ad-hoc, subjective fashion, on the whims of the reviewer. Finally, narrative reviews have historically suffered from a lack of adequate emphasis on inclusion and exclusion criteria for studies, resulting in too subjective an overview. Systematic reviews and meta-analysis, introduced in this paper, address many of the limitations of narrative review.

What are the Advantages of Systematic Review and Meta-Analysis? and What are their Differences?

Systematic review and meta-analysis (Littell et al. 2008) was developed to address the limitations of narrative review. A systematic review aims to provide a comprehensive literature search with pre-defined eligibility criteria whereas a meta-analysis combines and synthesizes findings with statistical models (Liberati et al. 2009). Therefore, systematic review focuses on minimizing bias in literature review by using clearly pre-defined criteria so that the literature search is replicable. On the other hand, meta-analysis statistically combines the effect sizes and models the effect sizes with study characteristics (Borenstein et al. 2009; Glass 1976; Schmidt and Hunter 2015).

Although systematic review and meta-analysis are usually conducted in the same study, there are subtle differences between them (e.g., Borenstein et al. 2009, pp. xxvii-xxviii). Some studies may use systematic and pre-defined eligibility criteria to do literature reviews. However, the researcher may choose not to conduct a meta-analysis for some reasons, for example, the extracted studies do not contain enough information to calculate the effect sizes for the meta-analysis. If the search strategy is flawed or the studies are biased, the findings of the meta-analysis are also biased. Therefore, it is advisable to combine both systematic review and meta-analysis in the same review process. On the other hand, some meta-analyses are not systematic reviews because the search process is not based on some pre-defined eligibility criteria. One example is when the researcher may want to combine the effect sizes based on several primary studies they have conducted. Thus, there is no systematic review in this meta-analysis.

There are several strengths of systematic review and meta-analysis. Systematic review ensures that the search strategies are replicable, that is, other researchers may extract the same studies by using the same searching procedures (Aytug et al. 2012). Replicability becomes more and more important in scientific research now (e.g., Francis 2013; Lindsay 2015; Lucas and Brent Donnellan 2013). Systematic review is the preferred way to do literature review for meta-analysis. It should be noted that systematic reviews may not be able to correct the problems in primary studies. However, a good systematic review may allow other researchers to replicate the same results in literature review and possibly in meta-analysis (Cooper et al. 2008).

In a meta-analysis, researchers focus not on statistical significance of individual studies, but on the magnitude of the effect or effect size. The effect sizes are weighed by the precision of the effect sizes (see Equations 1 and 2). Apart from the average effect, we may also estimate the heterogeneity of the effects, which indicates the consistency of the effect across studies. Researchers may also use study - level characteristics as moderators, to explain some of the variation in the effect size, thus providing a non-subjective way of interpreting differences in results among studies.

When should I as a Researcher Decide to Conduct (or not to Conduct) a Meta-Analysis?

Similar to the question "Should I conduct a study on topic X?", it is nearly impossible to set a simple rule to determine whether or not to conduct a meta-analysis on a particular topic. We suggest two questions that may help researchers to make their own decisions. The first question is whether there are enough primary studies for the meta-analysis. Meta-analysis is the analysis of primary studies. Although it is possible to conduct a meta-analysis with only two studies in principle (see the next section), conclusions based on only a few studies may not be well received as strong evidence by the research community. The availability of a small number of empirical studies also suggests that the field may not be mature enough to yield useful findings.

The second question is how important and pressing is the topic? If the topic is critical to human lives and the society, researchers may still want to conduct a meta-analysis even though there are not too many primary studies. The main objective of these meta-analyses is to increase precision of the estimate by pooling more data together. Although the findings may not be conclusive, they may still provide some insights on whether the current (limited) evidence is convergent.

It should be noted that the answers for these two questions may or may not be consistent. It may be the case that the topic is important and pressing but there are only limited empirical studies. It is always the researcher's duty to justify the contributions of the meta-analysis to the literature.

How Many Primary Research Articles are Required for a Meta-Analysis?

It is not easy to state the number of studies required for a meta-analysis. Factors affecting the decision may involve, say, discipline specific context, fixed- or random- effects models used in the analysis, population values of effect sizes, and other considerations. Koricheva et al. (2013) suggests that the range of number of studies is from 20 to 30 in ecology and evolution. Weare and Nind (2011) reviewed 52 systematic reviews and meta-analyses which met the selection criteria of being published in mental health. The range of studies was from 4 to 526 with a median of 35. Davey et al. (2011) reviewed 22,453 meta-analyses that met the selection criteria in one issue of the Cochrane Database of Systematic Reviews. They found that the median number of studies was 3 whereas the maximum number of studies was 294. It is evident that the number of studies varies across published meta-analyses.

One common approach to estimate the number of studies required is based on power analysis. Hedges and Pigott (2001, 2004, Pigott, 2012) provide detailed accounts of how to conduct power analysis in meta-analysis. Valentine et al. (2010) argues that it is still worthwhile to conduct a fixed-effects meta-analysis with only two studies. It is because the meta-analyzed estimates are still more precise than that of any individual study. On the other hand, more studies are required if a random-effects model (see Equation 3 for the definition) is used, otherwise, the estimated heterogeneity variance is not stable enough. Regardless of the actual numbers of studies included in the meta-analysis, researchers may need to justify whether the selected studies are a sufficient number for the meta-analysis.

How should the Articles Be Selected for a Meta-Analysis?

There are many databases such as PubMed, MEDLINE, Embase, Web of Science, Scopus, PsycINFO, and Google Scholar (Reed and Baxter 2009). It is difficult to say which databases should be used in the meta-analyses. Different disciplines may have their own preferences. In order to minimize selection bias, it is advisable to use more than one database. One approach is to go through the recent issues in the relevant journals. This strategy may give you some ideas on which databases are popular in your field. The literature searching procedure should also be as comprehensive as possible. Moreover, the search strategy should be well documented so that the search is replicable by other researchers.

A common question is whether unpublished studies such as dissertations, conferences, and unpublished papers, should be included in the meta-analyses. There are pros and cons of including unpublished studies. Ferguson and Brannick (2012) raises four potential concerns of including unpublished studies in the meta-analyses: (1) the unpublished studies may be of weaker methodology, (2) the included unpublished studies may be biased towards the authors conducting the meta-analyses because of the ease of availability of these studies, (3) searches for unpublished studies may favor established rather than non-established authors, and (4) the search for unpublished studies may also be biased. On the other hand, Rothstein and Bushman (2012) provides counter arguments for including unpublished studies. Specifically, they recommend: (1) studies may be excluded on the methodological rigorousness by using clearly define inclusion criteria rather than excluding all unpublished studies, (2) researchers may also try to contact the authors who have published in the topic to minimize potential selection bias, and (3) researchers should include unpublished studies and test whether study characteristics related to methodological quality moderate the effect sizes (also see Liberati et al. 2009).

What Are the Common Effect Sizes?

Meta-analysis is the statistical analysis of effect sizes. An effect size summarizes the effect of intervention, manipulation or observation of some phenomenon being studied. Effect sizes can be either unstandardized or standardized. Unstandardized effect sizes are appropriate if the effect sizes can be used to communicate or compare across studies. For example, blood pressure and heart rates are some good examples of unstandardized effects in medical science. If the meanings of the scales or measures are less clear, researchers tend to standardize them before conducting a meta-analysis. One basic requirement is that the effect size is directional to indicate the direction of treatment or association. Therefore, some common effect sizes, for example, R^2 in a multiple regression, η^2 and ω^2 in ANOVA, are usually not appropriate for a meta-analysis.

There are three common types of effect sizes (e.g., Borenstein et al. 2009; Cheung 2015a). The first one is based on some binary outcome, e.g., yes/no, failed/success, and life/death, in the form of a contingency table. This type of effect size is popular in medical and health sciences. Relative risk and odds ratio are some of these examples.

The second type is based on the mean differences. It is used to represent the treatment effect between an experimental and a control groups or the gender effect between males and females. If the scale is meaningful, for example, an IQ score, a raw mean difference or unstandardized mean difference may be used as the effect size. If the scale of the measure is not comparable across studies, we may calculate a standardized mean difference. For repeated measures designs, we may calculate effect sizes based on pre- and post-tests. If the effect sizes are chosen properly, the calculated effect sizes based on the between- and within-studies are comparable (Morris and DeShon 2002). Researchers may synthesize studies with both between- and within-studies.

The last type of effect size is based on the correlation coefficient. It is used to represent the association between two variables. Some researchers, for example, Schmidt and Hunter (2015), work directly with the correlation coefficient as various measurement and statistical artifacts can be corrected before conducting a meta-analysis. Other researchers, for example, Hedges and Olkin (1985), prefer to transform the correlation coefficient to Fisher's z score before applying a meta-analysis. This transformation helps to normalize the sampling distribution of the correlation when used as an effect size.

The choice of the type of effect size is usually based on the settings and research questions. If the researchers are working in the medical and health sciences, the typical outcome will be binary. If the research topic is related to experimental or between group comparisons, mean differences are usually used. If the studies are primarily based on observational studies, correlation coefficient is the obvious choice. That said, it is possible to convert the effect sizes among odds ratio, mean difference and correlation coefficient (Borenstein et al. 2009, Chapter 7). Researchers do not need to exclude studies because of the difference in the reported effect sizes.

Besides computing the effect sizes, researchers also need to compute the approximate sampling variances of the effect sizes. Most meta-analytic procedures use some form of weighted average to take the precision of the effect sizes into account. An effect size cannot be used in the analysis if the sampling variance of that effect size is missing. In this paper we use y_i and v_i to represent a generic effect size and its approximate sampling variance in the ith study. y_i can be either an odds ratio, standardized mean difference or correlation coefficient depending on the research question. Table 1, adopted and modified from Cheung et al. (2012), shows some common effect sizes and their approximate sampling variances. Interested readers may refer to Borenstein (2009) and Fleiss and Berlin (2009) for more details on how to calculate these effect sizes.

How to Choose between a Fixed- and a Random-Effects Models?

Before conducting a meta-analysis, researchers have to choose between a fixed- and a random-effects models (Borenstein et al. 2010). The fixed-effects model usually assume that the effect sizes are homogeneous across studies. The model for a fixed-effects univariate meta-analysis is

$$y_i = \beta_{\mathrm{F}} + e_i, \tag{1}$$

where β_{F} is the common effect for all k studies, and $\mathrm{Var}(e_i)=v_i$is the known sampling variance in the ith study. Thus, the fixed-effects model, also known as the common effect model, assumes that all the studies share the same population effect size β_{F}. The observed difference between the effect sizes of the studies is mainly due to the sampling error.

Researchers may calculate a Q statistic to test the assumption of homogeneity of the effect sizes. The Q statistic is defined as:

$$Q = \sum_{i=1}^{k} w_i \left(y_i - \beta_{\mathrm{F}}\right)^2, \tag{2}$$

where $\beta_{\mathrm{F}} = \sum_{i=1}^{k} (w_i y_i) / \sum_{i=1}^{k} w_i$ and $w_i = 1/v_i$ are the estimated common effect and the weight (and precision), respectively. Under the null hypothesis of the homogeneity of effect sizes, the Q statistic has an approximate chi-square distribution with (k-1) degrees of freedom. Statistically speaking, if the Q statistic is significant, the null hypothesis of the homogeneity of effect sizes is rejected at α=.05. However, it is generally found that the statistical power of the Q statistic is quite low in detecting the heterogeneity of effect sizes (e.g., Harwell 1997; Huedo-Medina et al. 2006; Viechtbauer 2007). On the other hand, the Q statistic is likely to be significant when there are lots of studies. Therefore, it is not advisable to choose between the fixed- versus the random-effects models by relying on the significance test on the Q statistic.

A random-effects model allows studies have their own population effect sizes. The model is

$$y_i = \beta_{\mathrm{R}} + u_i + e_i, \tag{3}$$

where β_{R} is the average population effect, and $\mathrm{Var}(u_i)=\tau^2$is the population heterogeneity variance that has to be estimated from the data. Thus, the random-effects model assumes that the observed difference on the sample effect sizes consists of two components: (a) the true differences among the population effect sizes, and (b) the sampling error.

Apart from interpreting the heterogeneity of the variance τ^2, researchers may also report the I^2 index, which can be interpreted as the proportion of the total variation of the effect size due to the between-study heterogeneity (Higgins and

Table 1 Common effect sizes and their sampling variances

Quantity of interest	Summary statistics required to compute the effect sizes	Effect size (y_i)	Approximate sampling variance (v_i)
Relative risk (*RR*) Odds ratio (*OR*)	*a*: frequency of success in Group 1 *b*: frequency of failure in Group 1 n_1 = a + b	$y_{RR} = \log\left(\frac{a^*n_2}{c^*n_1}\right)$	$v_{RR} = \frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}$
	c: frequency of success in Group 2 *d*: frequency of failure in Group 2 n_2 = c + d	$y_{\mathrm{OR}} = \log\left(\frac{a^*d}{b^*c}\right)$	$v_{\mathrm{OR}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$
Raw mean difference (RMD) Standardized mean difference (SMD)	X_1: Sample mean for Group 1 S_1^2: Sample variance for Group 1 n_1: Sample size for Group 1	$y_{RMD} = X_1 - X_2$	$v_{RMD} = S_{\text{pooled}}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$
	X_2: Sample mean for Group 2 S_2^2: Sample variance for Group 2 n_2: Sample size for Group 2 $S_{\text{pooled}}^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$y_{SMD} = \left(1 - \frac{3}{4(n_1+n_2)-9}\right)\frac{X_1-X_2}{S_{\text{pooled}}}$	$v_{SMD} = \frac{n_1+n_2}{n_1n_2} + \frac{y_{SMD}^2}{2(n_1+n_2)}$
Correlation (r) Fisher's *z* transformed score (z)	*r*: sample correlation coefficient *n*: Sample size	$y_r = r$	$v_r = \frac{\left(1-r^2\right)^2}{n-1}$
		$y_z = 0.5^*\log\left(\frac{1+r}{1-r}\right)$	$v_z = \frac{1}{n-3}$

Thompson 2002). Higgins et al. (2003) suggests that an I^2 of 25 %, 50 %, and 75 % respectively. It should be noted, however, that these suggestions are based on some meta-analyses in medical research. These guidelines may or may not be applicable to other fields. Another cautionary note is that I^2 is a relative measure of heterogeneity. It becomes larger when the sampling error gets smaller and vice versa. On the other hand, τ^2 is an absolute measure of heterogeneity that is theoretically free of influence from the sampling error. These two measures sometimes may not be consistent with each other. Researchers are recommended to report both measures so that readers are informed of the complete picture.

How to Conduct and Interpret Moderator Analysis?

When researchers suspect that some study characteristics may be used to explain the differences in the effect sizes, they may test these hypotheses by using a mixed-effects meta-analysis, which is also known as a meta-regression. Study characteristics, such as year of publication, type of interventions, mean age of the participants, proportion of female participants, countries of the participants can be used as moderators. The mixed-effects model is,

$$y_i = \beta_0 + \beta_1 x_1 + u_i + e_i, \tag{4}$$

where x_i is a study characteristic in the *i*th study, Var(u_i)=$\tau^2$is now the residual heterogeneity after controlling for x_i, β_0 and β_1 are the intercept and the regression coefficient, respectively. The results can be interpreted similarly to the familiar linear regression analysis with the exception that we are analyzing the study level phenomenon. This is quite crucial because we do not have the individual level data. Incorrectly inferring findings from the study level to the individual level is known as the ecological fallacy. The major concern is that the relationship at the group level may or may not hold at the individual level. Suppose we are using (i) the mean age of the participants and (ii) the proportion of female participants as moderators. If the estimated regression coefficients are both positive, they suggest that the effects are stronger in *studies* with older participants and with more female participants. Researchers should be cautioned not to interpret the findings as suggesting that the effects are stronger for older *participants* and *females*. A simple way to detect potential ecological fallacy is the change of referent from studies to individuals, for example, from mean age to age, from proportion of female participants to females.

What Is Publication Bias and how to Address it?

It is widely accepted that published studies may not be a representative sample of all possible effects (e.g., Francis 2013). For example, authors are less likely to complete their projects when most of their results are statistically non-significant (Easterbrook et al. 1991; Dickersin et al. 1987). Reviewers and editors may also be less likely to accept papers that show non-significant effects. Based on a survey of 80 authors of articles published in psychology or educational journals, Kupfersmid and Fiala (1991) found that 61 % of the 68 respondents believed that there is little chance of the manuscript being accepted if the research result is not statistically significant. Similar results were also demonstrated in several recent studies (see Song et al. 2010). This concern has recently been raised again as part of the discussion on whether psychology and research in general are replicable and reproducible (e.g., Lindsay 2015; Lucas and Brent Donnellan 2013; Open Science Collaboration 2012, 2015).

Since meta-analysis is largely based on accumulation of published studies, findings in meta-analyses may unavoidably be affected by publication bias. Several procedures, such as funnel plot, Egger's regression, trim and fill procedure, and fail-safe N, have been developed to address issues of publication bias. Because of the space limitation, readers may refer to, e.g., Rothstein et al. (2005) for the details. It should be noted these methods were developed based on different assumptions with the objectives to addressing different types of publication bias. Researchers are strongly advised to apply some of these methods to check whether the findings are compromised by the threat of publication bias.

What Are the Software Options Available?

There are many software packages available for meta-analysis (see Bax et al. 2007; Koricheva et al. 2013; Sterne et al. 2001; Wallace et al. 2009). Besides these software packages, there are also many R packages (Dewey 2015) implemented in the R statistical platform (R Development Core Team 2016). These software packages differ in terms of cost (commercial vs. free), user interface (graphical user interface vs. command line), types of effect sizes calculated, graphical output (with or without forest plot and other graphical output for meta-analysis), and statistical models for meta-analyses (simple analyses to sophisticated methods).

Instead of repeating these reviews, this paper uses artificial data to illustrate how to conduct common meta-analyses, e.g., fixed-, random- and mixed-effects models, in various software packages. The main objective is to provide a quick tour to users so that they may try different software packages and choose the one that fits their purposes. The software packages demonstrated here are the Comprehensive Meta-Analysis (Borenstein et al. 2005), SPSS macro (Lipsey and Wilson 2000), Stata (Palmer and Sterne 2016), Mplus (Cheung 2015a, Chapter 9), the metaSEM package (Cheung 2015b) implemented in the R statistical environment. Please refer to the supplementary materials or https://goo.gl/amYoGC for the analyses and outputs.

To make the example more relevant to the readers, we simulated data that examines the effect of schizophrenia on a dimension of cognitive ability, for example, on attention (see e.g., Fioravanti et al. 2005 for the relevant theoretical background). Our simulated meta-analysis uses fictitious studies that compared people with schizophrenia with a control group of healthy participants, in controlled studies on attention. The effect size in each study is a continuous measure, the standardized mean difference (SMD) or Hedges's *g*, between schizophrenia and control groups. People with schizophrenia are expected to perform less well than controls in measures of attention, hence the hypothesized effect size should be negative. To examine if study variables affected the differences in effect sizes among studies, we also included the mean age of participants in each study as a study-level variable. Please note that the artificial data only serve the purposes of illustrating the procedures. The results should not be interpreted substantively.

The following results are based on the metaSEM package in R. The results may be slightly different in other software packages because of the differences on the estimation methods. There are 50 studies in this illustration. The Q statistic is $\chi^2(49) = 85.71$, $p < .001$. The estimated heterogeneity variance is 0.0273 while the I^2 is .3955. These suggest that there is a moderate amount of heterogeneity. The between-study effect can explain about 40 % of the total variation whereas the other 60 % is due to the within-study variation. The average SMD and its 95 % confidence interval (CI) based on the random-effects model are −0.7287 (−0.8019; −0.6555). People with schizophrenia generally perform less than those in the controls. When the mean age of the participants is used as the moderator, the estimated regression coefficient and its 95 % CI are −0.0071 (−0.0128; −0.0014). Thus, the effect is stronger (more negative) for studies with older participants.

What Are the Common Pitfalls of Meta-Analysis and their Solutions?

Although meta-analysis has been generally accepted in many disciplines, it is not without criticisms. Borenstein et al. (2009, Chapter 43) provides a nice summary and responses to many of these criticisms. One common criticism is that meta-analysis is largely based on published studies. The presence of publication bias may affect the validity of the meta-analysis. It should also be noted that this criticism is not of meta-analysis per se. All reviews and comparisons involving published studies are also subject to publication bias. Meta-analysis, however, does provide various methods to address the potential issues of publication bias. In this sense, meta-analysis is considered as a good approach to address publication bias and replicability issues (e.g., Maxwell et al. 2015).

Another common criticism is that meta-analysis combines studies with different research objectives, designs, measures, and samples. This is known as the "apples and oranges" problem. If the studies are of low quality, the findings based on meta-analysis cannot be better than that. This is known as the "garbage in, garbage out" argument against meta-analysis. Eysenck (1978) once called the meta-analysis conducted by Smith and Glass (1977) as "an exercise in mega-silliness." As argued by Schmidt and Hunter (2015), synthesizing findings from different settings is in fact a strength of meta-analysis. The use of random- and mixed-effects model, as demonstrated in the illustrations, allows researchers to test the theory across different designs, measures and samples. More importantly, the variability of the effect sizes can be quantified and modelled by using study characteristics. Researchers may test how well the theory or treatment work under various conditions.

Besides these conceptual criticisms, another common criticism is that common meta-analytic procedures (e.g., Hedges and Olkin 1985) assume that the effect sizes are independent. It is rare that one study only reports one effect size. The extracted effect sizes are likely correlated or non-independent. If these effect sizes are treated as independent, the conclusions are likely incorrect. Several statistical procedures (e.g., Cheung 2013, 2014, 2015a) have been developed to address multiple and non-independent effect sizes. Researchers should apply some of these methods to handle the data properly.

Conclusions

Meta-analysis is a valuable research tool to synthesize research findings in many different disciplines. It is the preferable method to accumulate research findings in scientific research. Many of these limitations can be addressed by carefully formulating the systematic review and applying appropriate statistical meta-analytic models. This paper only provided a very brief introduction to some of these issues. There are many good and authoritative resources in meta-analysis (e.g., Borenstein et al. 2009; Cooper et al. 2009). It is our hope that this paper can further motivate readers to learn and apply meta-analysis in their research.

Acknowledgments Mike W.-L. Cheung was supported by the Academic Research Fund Tier 1 (FY2013-FRC5-002) from the Ministry of Education, Singapore. We would like to thank Maggie Chan for providing comments on an earlier version of this manuscript.

References

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15*(1), 103–133. doi:10.1177/1094428111403495.

Bax, L., Yu, L.-M., Ikeda, N., & Moons, K. G. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology, 7*(1), 40. doi:10.1186/1471-2288-7-40.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., & Rothstein, H. R. (2005). *Comprehensive meta-analysis (version 2)*. Englewood NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, U.K.; Hoboken: John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. doi:10.1002/jrsm.12.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions, 25*(1), 12–37. doi:10.1177/0163278702025001003.

Cheung, M. W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 429–454. doi:10.1080/10705511.2013.797827.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods, 19*(2), 211–229. doi:10.1037/a0032968.

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. Chichester, West Sussex: John Wiley & Sons, Inc..

Cheung, M. W.-L. (2015b). metaSEM: an R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*(1521). doi:10.3389/fpsyg.2014.01521.

Cheung, M. W.-L., Ho, R. C. M., Lim, Y., & Mak, A. (2012). Conducting a meta-analysis: Basics and good practices. *International Journal of Rheumatic Diseases*, *15*(2), 129–135. doi:10.1111/j.1756-185X.2012.01712.x

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 3–16). New York: Russell Sage Foundation.

Cooper, H., Maxwell, S., Stone, A., & Sher, K. (2008). Reporting standards for research in psychology why do we need them? What might they be? *American Psychologist, 63*(9), 839–851. doi:10.1037/0003-066X.63.9.839.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the Cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology, 11*, 160. doi:10.1186/1471-2288-11-160.

Dewey, M. (2015). CRAN task view: Meta-analysis. Retrieved from http://CRAN.R-project.org/view=MetaAnalysis

Dickersin, K., Chan, S., Chalmersx, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8*(4), 343–353. doi:10.1016/0197-2456(87)90155-3.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet, 337*(8746), 867–872. doi:10.1016/0140-6736(91)90201-Y.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128. doi:10.1037/a0024445.

Fioravanti, M., Carlone, O., Vitale, B., Cinti, M. E., & Clare, L. (2005). A meta-analysis of cognitive deficits in adults with a diagnosis of schizophrenia. *Neuropsychology Review, 15*(2), 73–95. doi:10.1007/s11065-005-6254-9.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–253). New York: Russell Sage Foundation.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153–169. doi:10.1016/j.jmp.2013.02.003.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. doi:10.2307/1174772.

Harwell, M. (1997). An empirical study of Hedge's homogeneity test. *Psychological Methods, 2*(2), 219–231. doi:10.1037/1082-989X.2.2.219.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 203–217. doi:10.1037/1082-989X.6.3.203.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445. doi:10.1037/1082-989X.9.4.426.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. doi:10.1002/sim.1186.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557–560.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis. *Q statistic or I2 index? Psychological Methods June 2006, 11*(2), 193–206.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage Foundation.

Koricheva, J., Gurevitch, J., & Mengersen, K. (Eds.) (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton: Princeton University Press.

Kupfersmid, J., & Fiala, M. (1991). A survey of attitudes and behaviors of authors who publish in psychology and education journals. *American Psychologist, 46*(3), 249–250. doi:10.1037/0003-066X.46.3.249.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ, 339*(jul21 1), b2700–b2700. doi:10.1136/bmj.b2700.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832. doi:10.1177/0956797615616374.

Lipsey, M. W., & Wilson, D. (2000). Practical meta-analysis. Sage Publications, Inc ,Thousand Oaks.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). Systematic reviews and meta-analysis. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195326543.001.0001/acprof-9780195326543

Lucas, R. E., & Brent Donnellan, M. (2013). Improving the replicability and reproducibility of research published in the journal of research in personality. *Journal of Research in Personality, 47*(4), 453–454. doi:10.1016/j.jrp.2013.05.002.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487–498. doi:10.1037/a0039400.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. doi:10.1037/1082-989X.7.1.105.

National Research Council (1992). *Combining information: statistical issues and opportunities for research*. Washington, DC: National Academies Press.

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *JRSM, 100*(12), 579–582. doi:10.1258/jrsm.100.12.579.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660. doi:10.1177/1745691612462588.

Open Science Collaboration (2015). Estimating the reproducibility of psychological. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716.

Palmer, T. M., & Sterne, J. A. C. (Eds.) (2016). *Meta-analysis: an updated collection from the Stata journal* (2nd ed.). College Station: Stata Press.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *BMJ, 2*(2288), 1243–1246. doi:10.1136/bmj.2.2288.1243.

Pigott, T. D. (2012). Advances in meta-analysis. New York: Springer. Retrieved from http://www.springer.com.libproxy1.nus.edu.sg/statistics/social+sciences+%26+law/book/978-1-4614-2277-8

R Development Core Team. (2016). *R: a language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 73–101). New York: Russell Sage Foundation.

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: comment on Ferguson and Brannick (2012). *Psychological Methods, 17*(1), 129–136. doi:10.1037/a0027128.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester: John Wiley and Sons.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Shadish, W. R. (2015). Introduction to the special issue on the origins of modern meta-analysis. *Research Synthesis Methods, 6*(3), 219–220. doi:10.1002/jrsm.1148.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752–760.

Song, F., Parekh, S., Hooper, L., Loke, Y., Ryder, J., Sutton, A., et al. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment, 14*(8). doi:10.3310/hta14080.

Sterne, J. A. C., Egger, M., & Sutton, A. J. (2001). Meta-analysis software. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic Reviews in Health Care*: *Meta-Analysis in Context* (pp. 336–346). London: BMJ Publishing Group. Retrieved from doi:10.1002/9780470693926.ch17/summary

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215–247. doi:10.3102/1076998609346961.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60*(1), 29–60. doi:10.1348/000711005X64042.

Wallace, B. C., Schmid, C. H., Lau, J., & Trikalinos, T. A. (2009). Meta-analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology, 9*(1), 80. doi:10.1186/1471-2288-9-80.

Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say? *Health Promotion International, 26*(suppl 1), i29–i69. doi:10.1093/heapro/dar075.