What's fruit got to do with meta-analysis?

So far on the meta-evidence blog, we have introduced readers to some of those tricky issues that we run into when synthesising evidence on a particular topic. This week I attempt to disentangle some misconceptions around heterogeneity in meta-analysis and, as always, I will provide some top tips.

The classic apples and oranges criticism against meta-analysis has been stubbornly persistent for almost four decades and arguably 'fruitful' in causing trepidation for those who wish to conduct evidence synthesis. In a 1984 paper by Eysenck, ominously titled 'Meta-analysis: An abuse of research integration', he concluded:

> Adding apples and oranges may be a a pastime for children learning to count, but unless we are willing to disregard the differences between these two kinds of fruit, the result will be meaningless.

This argument offered by Eysenck is related to the assumed heterogeneity created when combining dissimilar studies, with some being apples and others being oranges. To understand this argument fully, I must now explain what heterogeneity is.

What is Heterogeneity?

Heterogeneity and Homogeneity are common terms relating to the uniformity of a given 'thing'. When meta-analysts talk about heterogeneity, they are referring to the variation across studies or results. Conversely, homogeneity refers to the similarities across studies or results.

We become interested in how much heterogeneity exists in a meta-analysis as this may signal that the results of the intervention or process under consideration may not easily generalise to other

contexts or populations. As Emily Tanner Smith discussed in an earlier post, the goal of modern meta-analysis is not simply to report the mean effect size across primary studies but to recognise how the effect sizes in the primary studies are dispersed around the mean. When we take the time to investigate heterogeneity properly, we can identify which factors influence results, which is an extremely valuable contribution to make.

When we take the time to investigate heterogeneity properly, we can identify which factors influence results, which is an extremely valuable contribution to make.

Factors which may increase variation across studies and results include methodological differences, statistical differences, and contextual differences. We will briefly consider each of these factors.

Methodological differences which lead to heterogeneity in meta-analysis may arise through the grouping of dissimilar study designs, such as the combination of cluster-randomised and quasi-randomised trials. Similarly, this may be the case when studies at a high risk of bias are mixed with studies at a low risk of bias. The reason why combining studies of varying quality and design leads to heterogeneity is due to differences in effect sizes.

Contextual differences. In health-related reviews such as those conducted though the Cochrane Collaboration, this contextual difference is most often referred to as clinical heterogeneity. Contextual difference or clinical heterogeneity refer to differences between the study outcomes and real-world outcomes. This heterogeneity may arise when the populations or interventions tested are nonuniform.

Statistical differences. We have described how heterogeneity can appear through both methodological and contextual factors. Statistical heterogeneity is the assumption that due to methodological and/or contextual differences then results vary more than is due to random error or chance. Statistical heterogeneity is the differences between primary studies, not due to chance. Heterogeneity will always be present, but it is important to understand the amount that exists.

Calculating heterogeneity

Statistical heterogeneity can be checked in a number of ways (Higgins, Thompson, Deeks, & Altman, 2003). First, visually using forest plots and checking for overlap of confidence intervals. Second, using tests such as the Cochran Q test (Chi-Square or c^2), percentage of total variation across studies (I^2) and the Tau-squared statistic (τ^2 or Tau^2).

When using the Cochran Q test, authors often agree the presence of heterogeneity when $p<0.1$. This figure may be chosen as it counterbalances the relatively low power of the test. In cases where there are a large number of included studies, Q is expected to be highly significant.

The I^2 test represents the total variation across studies and is unlike the Q test in that it is independent from the number of studies; instead I^2 is based on treatment effect and outcomes. The following equation from the Cochrane handbook shows how I^2 and Q are interrelated:

$$I^2 = 100\% \times (Q-df)/Q$$

I^2 ranges from 0-100% with 0% representing total absence of observed heterogeneity. The impact of heterogeneity was determined as low (25%), medium (50%) or high (75%) (Higgins et al., 2003). Finally, $\tau 2$ observes statistically significant heterogeneity when >1. Tau is the difference between total observed variance (Q) and within-studies variance (Higgins, Thompson, Deeks, & Altman, 2003).

If substantial heterogeneity is detected, many reviewers decide not to combine the effect sizes or present a synthesis of the findings. Others, however, investigate which study characteristics might be influencing the level of heterogeneity through techniques known as moderator analysis (Maynard, McCrea, Pigott, & Kelly, 2013).

Addressing heterogeneity

[caption id="attachment_801" align="alignleft" width="200"]  By investigating heterogeneity, we can identify which factors influence results[/caption]

Moderator analysis is where explanations for heterogeneity are explored through analysis of certain characteristics of the study. It can be handled in a way that is analogous to the one-way analysis of variance (ANOVA) and known as Subgroup analysis; or analogous to linear regression in primary research, known as meta-regression. The decision of which type of moderator analysis to use will often depend on the type of characteristics available.

Moderator analysis is where explanations for heterogeneity are explored through analysis of certain characteristics of the study.

A subgroup analysis will calculate the standardised mean difference (SMD) within each subgroup and then compare effectiveness and heterogeneity with the other subgroups in the category. Subgroup analysis will present details about the variance within the subgroups (Q_w) which is unexplained, and the variance between the subgroups (Q_b), and whether those differences are statistically significant.

Meta-regression differs slightly from subgroup analysis as the technique allows multiple continuous variables such as mean age or Risk of Bias score to be investigated simultaneously, as well as categorical variables if entered into the model as a series of dummy variables. In a meta-regression, the outcome variable is the SMD and the characteristics extracted are the predictors or criterion variables.

A meta-regression analysis can be represented by a simple scatter plot, with the variable of interest presented along the x-axis, and the SMD along the y-axis. The statistical software package, R, also allows the precision of each primary research to be proportional to the size of the plotting symbols provided. In addition to testing the statistical significance of the potential moderators on the SMD, it is also important to test the fit of the model using the coefficient of determination, also known as

the R^2 index. This index calculates the proportion of the variance of the SMD that is explained by the meta-regression model and covariates chosen to test.

It is important to understand that both types of moderator analyses are exploratory and should never be implemented to test hypotheses.

It is important to understand that both types of moderator analyses are exploratory and should never be implemented to test hypotheses. Even if the meta-analysis contains only random and quasi-random trials, the studies involved in these moderator analyses have not been randomised, they are observational in nature and at a higher risk of bias. Additionally, these type of analyses generally have lower power due to missing data in the primary research, there is an increased risk of presenting incorrect results which appear simply through chance (false positive conclusion), and potential for various biases (Borenstein, Hedges, Higgins, & Rothstein, 2009; Higgins & Green, 2011).

Heterogeneity and Statistical models

When primary studies are synthesised in a meta-analysis, they are usually combined using one of two statistical models: either a Fixed Effect Model (FEM) or a Random Effects Model (REM). The underlying assumption of a FEM is that there is one true effect size which underlies all the primary research included in the meta-analysis and that any differences observed between these studies is within study variance which is due merely to chance. The REM, in comparison, accepts two main differences among primary studies, the first is within study variance, and the second is between study variance. This between study variance, or heterogeneity, allows that difference between studies is always present due to important differences such as populations, settings, or progression of time (Borenstein, Hedges, Higgins, & Rothstein, 2009).

If all studies were equally accurate, reviewers could straightforwardly compute an average of each studies effect size, as this is highly unlikely, weights should be assigned. Weights allow those studies that are more precise estimations of the effect to contribute extra information. The choice of FEM or REM directly influences how weights are assigned to the individual studies (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Weights allow those studies that are more precise estimations of the effect to contribute extra information.

In a FEM, larger studies are assigned the most weight as it is assumed that it is a better representation of the true effect size, alternatively, in the REM, the aim is not to measure the true effect size, but to estimate the mean of the effect distribution. Since the underlying assumption is that each study provides unique information from a different sample, the REM does not assign most weight to a large study, and least to a small one (as would be done in FEM), but instead ensures that all studies are represented by their corresponding weights in the combined SMD. The confidence intervals are much wider in a REM than a FEM; this is due to the between study variance assumed by the REM, and so is a more conservative estimate of SMD (Raudenbush, 1994).

In Conclusion

We have covered what heterogeneity is and how to appropriately calculate and address it in meta-analysis. Through this deeper understanding, we start to realise that by investigating the sources of variance we are advantaged with the ability to explain potential differences in effect sizes. This is particularly important for Campbell review authors where many of us work in fields which accept and embrace complex systems perspectives and are naturally moving from a ‘what works’ linear cause and effect towards an understanding of what works, for whom, and in what circumstances?

I look forward to the day where we can all respond to the antiquated apples and oranges critique with the wit and confidence of Gene Glass and exclaim: "Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial."

Top Tips

[caption id="attachment_799" align="alignright" width="300"]

It's unlikely that primary studies in Campbell reviews are drawn from homogeneous samples[/caption]

Top tip one: The statistical model you choose should be based on the heterogeneity assumed rather than heterogeneity observed. This means that we should choose the model from the outset based on the sample of studies we have located and never based on a statistical test for heterogeneity. In most Campbell reviews, we are synthesising studies drawn from diverse populations; this would mean that a REM model is chosen.

Top tip two: Just as we address heterogeneity, statistically homogeneous results should be investigated by researchers who must discuss how this finding can be applied to real world contexts. For example, a homogeneous sample of genetically identical mice may produce zero variance across results, these results will not generalise to a population of humans.

References

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons

Eysenck, H. J. (1984). Meta-analysis: An abuse of research integration. *The journal of special education,* 18(1), 41-59.

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., . . . Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj, 343*, d5928. doi:10.1136/bmj.d5928

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal,* 327(7414), 557. doi:10.1136/bmj.327.7414.557

Maynard, B. R., McCrea, K. T., Pigott, T. D., & Kelly, M. S. (2013). Indicated truancy interventions for chronic truant students: A Campbell systematic review. *Research on Social Work Practice,* 23(1), 5-21. doi:10.1177/1049731512457207

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (Vol. 421). New York: Russell Sage Foundation.

Blog post written by Ciara Keenan