Total recall: Should meta-analyses have expiration dates?

Hannah R. Rothstein
Baruch College
City University of New York

WHY am I giving this talk?

- There are many excellent meta-analyses out there, but also many substandard ones.
- Some are substandard because they were done poorly to begin with.
- Others are substandard because they were done years ago and no longer measure up to current standards.
- An uncritical acceptance of all meta-analytic results as equally valid poses a danger to the credibility of the entire meta-analytic enterprise.

Why am **I** giving this talk?

- Meta-analysis and I got our start in research at about the same time.
- 1977, UMD Conference on Moderator Research- F Schmidt was presenting some of his & J Hunter's early ideas.
- 1986, ETS Conference on Test Validity--L Hedges & F Schmidt were both presenting their work.
- I've been an eyewitness to a scientific revolution and to the evolution and maturation of this important method of reviewing and synthesizing research.

Agenda

- Very brief history of meta-analysis.
- Differentiation of the terms systematic review and meta-analysis.
- Overview of key changes in best practices.
- Implications of these changes for the validity of early (and not so early) meta-analyses, and the research built on them.
- Suggestions for how to proceed.

Emergence of meta-analysis

- Quantitative combination of research findings go back over 100 years, but “modern” research synthesis is only about 35 years old.
- The term “meta-analysis” was introduced in 1976.
- The idea that reviews of scientific research literature should be conducted according to scientific principles was introduced at about the same time, although the term “systematic review” was adopted only in the 1990s.
- In a relatively short period of time, these ideas and methods have achieved enormous uptake.

Key definitions

- SYSTEMATIC REVIEW
  - The application of strategies that limit bias in the
    - Specification of a research question
    - Location of studies
    - Screening of studies for inclusion/exclusion
    - Coding of data from included studies
    - Synthesis of studies(which may or may not be meta-analytic).
- META-ANALYSIS
  - The statistical synthesis of the data from separate but similar (comparable) studies, leading to a quantitative summary of the combined results.

Distinct but complementary

- Meta-analysis emphasizes appropriate *analysis and interpretation* of the data
- Systematic review emphasizes the avoidance of biases *other than* in the integration of results.

Fundamental Synthesis Principles

- Explicit (e.g. in its statement of objectives, materials and methods)
- Systematic (e.g. in its identification of literature)
- Transparent (e.g. in its criteria and decisions)
- Reproducible (e.g. in its methodology and conclusions)
- Unbiased

Can’t have one without the other?

- Adoption of systematic review techniques and of meta-analysis did not take place in tandem.
- Some disciplines emphasized SR, others emphasized MA.
- Joint use as “best practice” in research synthesis is relatively recent, and uneven across disciplines.

We have reason to question

- A systematic review without a methodologically defensible means of synthesizing the data, OR a meta-analysis that is not preceded by a systematic review
  - Early meta-analyses were often conducted without a systematic review.
  - Some SRs that could have used MA didn't

Changing standards

- Methods for conducting systematic reviews and for conducting meta-analyses have evolved considerably since their inception;
- What was considered a “state of the art” systematic review OR meta-analysis even a few years ago may seem regrettably inadequate today.
- Let’s consider this in more detail

Best practices have changed re:

- When a meta-analysis is appropriate
- What to do when it is not
- Literature search
- Evaluation of retrieved studies
- Model choice
- Assessment of heterogeneity
- Publication bias & other sensitivity analyses
- Reporting practices

Let's review these one by one with examples

When is meta-analysis appropriate (1)

- The two main reasons given for not using meta-analysis to synthesize a literature are:
  - Too few studies
  - The studies are “too different” from each other to combine.
- There are differences across disciplines—social sciences versus health sciences
- But also inconsistency within disciplines

When is meta-analysis appropriate (2)

- Over time, the use of meta-analysis to summarize the data in a systematic review has increased in all disciplines
- Much of this has been due to an increased understanding of the limitations of popular alternatives such as
  - "Cognitive" algebra
  - Vote counting
- Let's look at an example

When is meta-analysis appropriate
Studies were “too different”

- Van Sluijs, McMinn & Griffin (BMJ, **2007**). Systematic review of interventions to promote physical activity in children and adolescents
- “We thought a formal meta-analysis inappropriate owing to the heterogeneity of the interventions, settings, participants, and outcome measures. Instead we used a rating system of levels of evidence to draw conclusions on effectiveness…”(p.5)

When is meta-analysis appropriate Studies were “too different” (2)

- Here’s their system
- “Conclusions were drawn on the basis of the consistency of results of studies with the highest available level of quality. If at least two thirds (66.6%) of the relevant studies were reported to have significant results in the same direction then we considered the overall results to be consistent.”
- This is vote counting

When is meta-analysis appropriate? What about vote counting?

- If a study found a statistically significant positive effect, a vote is recorded in column A
- If a study found a statistically significant negative effect, a vote is recorded in column B
- If a study failed to reject the null hypothesis, a vote is recorded in column C

When is meta-analysis appropriate? What about vote counting? (2)

- Vote counting ignores Type II error
  - Since many studies are conducted with low statistical power, a likely results is that the NS column will have the most votes
  - NS is incorrectly interpreted as “No effect” & wrong conclusion is reached
  - Power of the VC tends toward zero as number of studies in the analysis increases
- Vote counting gives Large N and small N studies the same weight
- Vote counting does not produce an effect size estimate

When is meta-analysis appropriate? Current view

- Short answer to "how many studies":
  - TWO
- Longer answer:
  - Are you interested in the overall effect, or in moderators of effect….?
  - What are you planning to do instead?
- Rebuttal to "data are too heterogeneous":
  - The breadth of the research question should determine the diversity of studies to be included
  - (We'll get back to heterogeneity later)

What to do if you can't meta-analyze?

- The next two slides show
  - A traditional narrative summary of three fictitious studies in the same review
  - A more informative and systematic alternative

Mays et al. (1999) 100 students were randomly assigned to participate in a prevention program or to be on a wait-list. Program effects were assessed via the Beck Depression Inventory. There was no significant difference between the groups.

Mantle et al. (2000) 20 students attending a local prevention program were compared to 20 ‘‘matched controls’’ on a previously published scale tapping attributional style. It was unclear how matching was implemented. There were no significant differences between the groups.

Snider et al. (2001) 60 students attending a prevention program at a local Boys & Girls Club were compared to 60 controls. The groups were matched on multiple measures of psychological functioning. Results revealed a positive effect, with students attending the prevention program performing better on a locally developed measure of depressive symptoms than students not attending the program.

Study/ Year	Treated N	Control N	Assignment Mechanism	Matching Used?	Outcome	ES	CI
Mays et al 1999	50	50	Random	No	Beck Depression Inventory	.15	+/- 0.34
Mantle et al 2000	20	20	Nonrandom	Yes, but variables unknown	Published attributional style scale	.30	+/- 0.62
Snider et al 2001	60	60	Nonrandom	Yes, multiple measures of Psych. Fcning.	Ad Hoc measure of depressive symptoms	.37	+/- .36

Current view: How to present data if you don’t combine meta-analytically

- Include text and/or tables for each study. describing the nature of the sample intervention (or predictor) outcomes (criterion variables) and research design, an estimate of the effect size, and 95% CI.
- Include a forest plot (graph) with a line for each study, but without the summary or overall effect line.

Literature Search

- Clear disciplinary differences
- But also inconsistent practices within disciplines
- Changes over time
  - Standards
  - Technology
  - Documentation (searches, lists of included studies)
  - Concern with publication bias
    - (more on this one later)

Literature Search: What's missing?

- "In addition to reviewing the published literature, we contacted numerous state and local jurisdictions and consultants and two law enforcement-related professional associations in an attempt to locate unpublished criterion-related validity studies of law enforcement personnel."
  - (Hirsh, Northrop & Schmidt, 1986)

Literature Search: What's missing in this one?

- "We conducted a search of the OCB literature by using a number of online databases (e.g., Web of Science, PsycINFO) as well as by examining the reference lists of previous reviews".
  - (Hoffman et al., JAP 2007)

Literature Search: Current Expectations

- To identify relevant studies, we first searched several computerized databases (ABI/INFORM, Business Source Complete, JSTOR, ProQuest, PsycARTICLES) using the search terms *turnover, quit, discharge, layoff, dismissal,* and *termination* in combination with the terms *organizational, collective, unit, proportion, rate,* and *ratio.* No limitations were placed on the year of publication. Second, a manual search of articles published in *Academy of Management Journal, Journal of Applied Psychology, Journal of Management,* and *Personnel Psychology* was conducted from the year 2000 forward. Third, citation searches were conducted for articles referencing seminal studies addressing collective turnover (e.g., Batt, 2002; Shaw et al., 1998; Staw, 1980). Fourth, we scanned reference lists of relevant articles. Fifth, to help mitigate possible publication bias, a computerized search of conference programs/proceedings was conducted for both the Academy of Management Annual Meeting and the Society for Industrial and Organizational Psychology Conference from the year 2007 forward due to the availability of electronic databases for this period. In the same vein, the ProQuest Dissertations and Theses database was searched using the aforementioned search terms. - (Heavey, Holwerda & Hausknecht, 2013 JAP)

But…

- Heavey, Holwerda & Hausknecht (2013) don’t do any publication bias analysis.
- Neither do Hoffman et al. (2007).
- Hirsh, Northrop and Schmidt (1986) did file-drawer analysis-Rosenthal’s version and Orwin’s.

Evaluating studies for inclusion

- In the past, often there were no criteria.
- Currently, we expect a priori inclusion criteria that assure us that the studies included are suitable to answer the research question:
- Study designs
- Quality
- Risk of bias
- Plus evidence of reliability of ratings and judgments

Model choice (1)

- Fixed or random and how to choose
  - Inappropriate use of the fixed model may overweight large studies, underestimates the SE produces CIs that are too narrow; and ignores heterogeneity
  - Until about 2003, the majority of psychology meta-analyses were FE.
  - Hunter, Oh & Hayes (2008) found that 76% of the meta-analyses published in Psychological Bulletin 1978 -2006 used only FE.

Model Choice (2)

- Schmidt et al. (2008) used RE to reanalyze 68 analyses from five large meta-analytic studies which had used FE.
- Key findings
  - published FE confidence intervals (CIs) around mean effect sizes were on average 52% narrower than their actual width,
  - nominal 95% FE CIs were found to be on average 56% CIs

Model Choice & Moderators (3)

- Back in the day, both Hedges (1982) and Rosenthal & Rubin (1982) recommended the use of a chi-square test to decide whether heterogeneity was "significant"
  - This also affected model choice (FE vs RE)
- Hunter & Schmidt originally proposed a 75 % that variance unexplained by artifacts should be more than 25% of the total observed variance in order for the unexplained variance to be considered significant
- These are no longer recommend, but many have used one or the other, and some researchers still use the chi-square test to determine model choice.

Forcing an FE model on the data

- Back in the day, Hedges & Olkin (1985) recommended the removal of outliers to achieve homogeneity (that is, to attain a non-significant value for the chi-square test.
- Hedges (1987) suggested that up to 20% of the ES could be removed as aberrant values.
- Hedges changed his view. This is now discouraged, but many researchers have used, and some still use this procedure.
- Let's look at 2 that IOOB rely on

Example 1: Outliers Removed

- Eagly & Karau (1991, JPSP). Gender and leadership.

"We attained homogeneity among the effect sizes by identifying outliers and sequentially removing those that reduced the homogeneity statistic by the largest amount"

Example 2: outliers removed

- Deci, Koestner & Ryan (1999, PB)
  - Effects of extrinsic rewards on intrinsic motivation

college students. The analysis revealed a significant difference between the two age groups, $Q_b(1) = 6.76$, $p < .01$. The set of 43 studies with children yielded a $d = -0.46$ (CI = −0.55, −0.37), but it was heterogeneous, $Q_w(42) = 118.85$, $p < .001$, so 4 outliers were removed (Boggiano, Ruble, & Pittman, 1982; Danner & Lonky, 1981, Exp. 2; Morgan, 1983, Exp. 1; Swann & Pittman, 1977, Exp. 2). The resulting set of 39 effect sizes yielded a composite $d = -0.43$ (CI = −0.53, −0.34) and was homogeneous, $Q_w(38) = 51.09$, *ns*. The 12 college-student studies, which

Assessment of heterogeneity

- There has been a lot of confusion about the various indices of heterogeneity and what information they provide.
- When heterogeneity statistics are presented in the results section of a meta-analysis, heterogeneity often isn’t considered in the discussion where the focus, is on the overall mean effect, rather than true dispersion of effects.

Publication Bias (1)

- Studies with significant, “positive”, results are easier to find than those with non-significant or 'negative' results; the same is true for outcomes within studies.
- Over-representation of “positive” studies in a meta-analysis can lead to an overestimate of effects, and an underestimate of heterogeneity.
- Meta-analyses in the Organizational Sciences often lack a publication bias analysis, or use outmoded methods.

Publication Bias (2)

- Aytug, Rothstein, Kern & Zhou (2011) found that only 18% of the studies in their sample of IOOB meta-analyses reported whether or not they conducted some type of publication bias analysis. For the studies reporting publication bias analyses, the most popular methods were Fail-safe N (49%) and comparison by study source (40%).
- Kepes et al. (2012, ORM) offer alternative means of assessing publication bias.

Sensitivity Analyses

- The human element is always there-
  - There are choices and judgments to make, and they have a bearing on the outcome.
- The robustness of the findings to different assumptions and decisions should be examined through sensitivity analysis.
- Aytug et al. found that only 16% of studies reported conducting any sensitivity analysis.

Reporting

- Guidelines in different fields emerge at different times; dissemination & uptake variable
  - Rothstein & McDaniel (1989)- widely ignored
  - QUOROM (1999) succeeded by PRISMA (2009) & its extensions (2012-2013)
  - MOOSE (2000)
  - MARS (2008)
  - RAMESES (realist and narrative syntheses, 2013)

Reporting of IOOB meta-analyses

- Aytug, et al.
  - On average, the meta-analyses in the sample provided 52.8% of the information needed to replicate the meta-analysis or to assess its validity and 67.6% of the information considered to be most important
  - More recently published meta-analyses exhibited somewhat more transparent reporting practices than older ones did
  - The correlation between transparency and number of citations was negative, but not significant (r = -.09)

Reliability and validity of meta-analyses

- My concern is the replicability of the results and conclusions of meta-analyses and their vulnerability to threats to validity.
- NB: Not a criticism of authors ,journal reviewers or editors (well, not of individual authors, reviewers or editors, but there is a systemic peer review problem which adds to the trouble

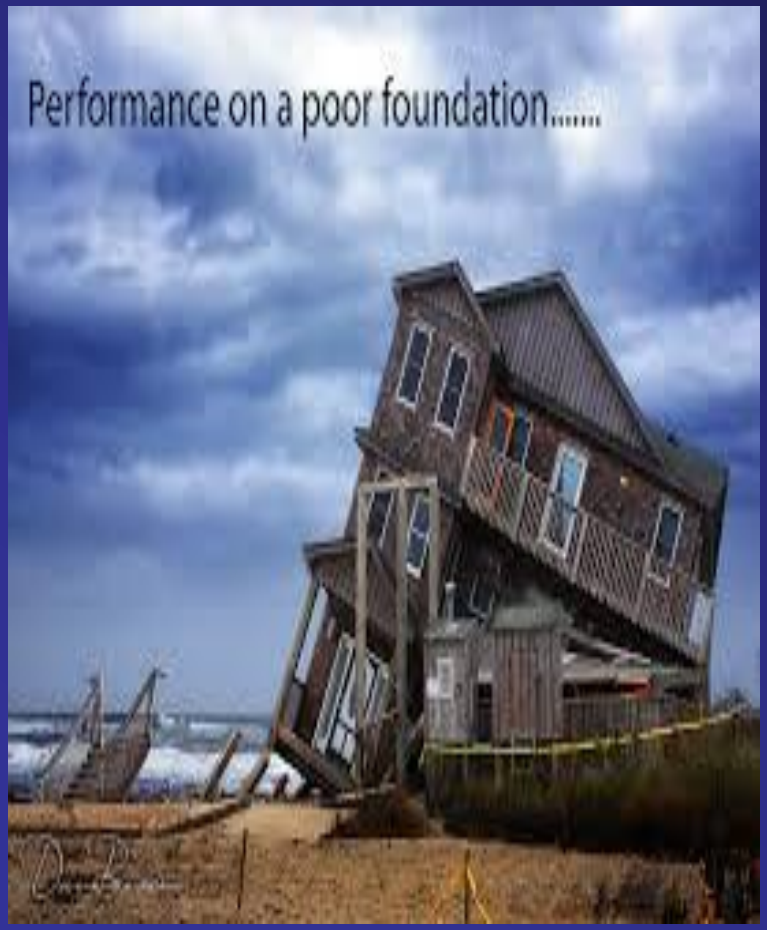
The reviewing crisis adds to the problem

- Lack of rewards for being a peer reviewer
  - Overburdened competent reviewers
    - Increased conflicts of interest?
  - More incompetent reviews
  - Reviewers often unfamiliar with policies/guidelines of journals for which they review-
- More bad papers out there, including bad meta-analyses
- Quite a few journals do not seem to require a meta-analysis expert to review meta-analyses.

Re-analysis and updating are needed

- Researchers revisit the adequacy of primary studies that were done according to outmoded standards**– we should treat meta-analyses the same way**.
- I propose that all meta-analyses done before 2010 be reviewed, and if found to be problematic, they should be updated or declared to be of uncertain validity
- If this is too burdensome, we can start with the most highly cited meta-analyses.

Which future do we choose?



Keep
calm
and
thank you
for listening

Articles Cited

Aytug Z, Rothstein HR, Zhou, W, Kern M. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analysis *Org Rsch Methods,* 15, 103-133.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). The undermining effect is a reality after all: Extrinsic rewards, task interest, and self-determination. *Psych Bull*, *125*, 692-700.

Eagly AH, Karau,S J (1991). Gender and the emergence of leaders. *J Pers Soc Psych,* 60, 685-710

Heavey A, Holwerda J, Hausknecht JP (2013). Causes and consequences of collective turnover: A meta-analytic review. *Jnl Appl Psych,* 98 (3), 412-453.

Hirsh HR, Northrop L, Schmidt FL (1986). Validity generalization results for law enforcement occupations. *Pers Psych*

Hoffman BJ, Blair CA, Meriac JP, Woehr DJ (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Jnl Appl Psych*, 92(2), 555-566.

Kepes, S, Banks GC, McDaniel, M, & Whetzel, DL (2012). Publication bias in the organizational sciences. *Org Rsch Methods,* 15,) 624-662

Schmidt FL, Oh IS, Hayes TL (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *Bri Jnl Math Stat Psych*, 62, 97–128.

Van Sluijs E, McMinn A & Griffin SJ (2007). Effectiveness of interventions to promote physical activity in children and adolescents: systematic review of controlled trials. *BMJ* , 335, 703