


How Big is the Effect? A Brief Intro to Meta-Analysis

Rob Cribbie
Quantitative Methods
(previously Social-Personality)

What is Meta-Analysis?

- ▶ The statistical summarization of the effects from a set of studies investigating the same research question
 - However, the term 'meta-analysis' often also applies to the entire process of generating a research question, finding studies that investigate the research question, extracting the necessary info from the studies, and combining the results from the related studies

Why Perform a Meta-Analysis?

- ▶ A single study cannot be used to definitively quantify the magnitude of an effect
 - Results (effects) vary from study to study due to sampling error, nature of the population, methodological procedures, etc.
 - ▶ Unsystematic or narrative reviews of the literature are often extremely biased from both the perspective of the methods and the researcher
 - E.g., the researcher usually has an a priori inclination regarding the conclusions of the study
- 

Meta-analysis in Research

▶ Publications

- Journals, as well as other researchers, encourage meta-analyses
- Meta-analyses provide a great starting point for research as they help contextualize a new study

▶ New Research

- Meta-analyses can be used as a tool to help researchers avoid *recreating the wheel*, or to find promising research areas by investigating past studies

▶ Grant Applications

- Meta-analyses are highly regarding in grant applications, as they contextualize the proposed research and reduce the likelihood that resources are wasted on effects known to be null

Systematic Review

- ▶ In some instances “systematic review” and “meta-analysis” are used interchangeably, whereas in other instances the term *systematic review* refers to the procedures used to collect the studies of interest (i.e., those to be combined), and *meta-analysis* refers to the statistical combination of the effects from these studies
 - Systematic Review
 - A review of studies addressing a research question that is conducted according to clearly stated methods

Some History from Psychology

- 1952: Hans Eysenck concluded that there were no favorable effects of psychotherapy, starting a raging debate
 - 20 years of evaluation research and hundreds of studies failed to resolve the debate
- 1978: To prove Eysenck wrong, Gene Glass statistically aggregated the findings of 375 psychotherapy outcome studies
 - Glass concluded that psychotherapy did indeed work
- Glass called his method “meta-analysis”

The Emergence of Meta-analysis

- Ideas behind meta-analysis predate Glass' work by several decades
- Karl Pearson (1904)
 - Averaged correlations for studies of the effectiveness of inoculation for typhoid fever
- R. A. Fisher (1944)
 - We can combine the results of several studies to get an appreciation for the probability associated with the aggregated data
 - Dealt primarily with combining p -values
- The start of the idea of *cumulating probability values*, although not specifically focused on effect sizes

The Emergence of Meta-analysis

- W. G. Cochran (1953)
 - Discussed a method for averaging means across independent studies
 - Cochran was responsible for much of the statistical foundation that modern meta-analysis is built upon
- Cochrane Collaboration
 - A group of researchers from around the world that conduct systematic reviews of health-care interventions and diagnostic tests and publish them in the Cochrane Library
 - <https://canada.cochrane.org/>

The Logic of Meta-analysis

- Traditional methods of review focus on statistical significance testing
 - E.g., the effect was statistically significant in 4 out of 7 studies
 - However, we know that NHST is highly related to sample size, focuses on dichotomous decisions, etc.
- Meta-analysis focuses on the *direction* and *magnitude* of the effects across studies, not statistical significance
 - Direction and magnitude are represented by the effect size

When Can You Do Meta-analysis?

- Studies are empirical, not theoretical
- Results are quantitative, not qualitative
- Studies examine the same research question
- Results can be quantified in a comparable statistical form
 - i.e., effect size

Research Questions Amenable to Meta-analysis

- Central tendency research (e.g., means)
 - Pre-post contrasts
 - Group contrasts
 - Experimentally created groups
 - E.g., change in perfectionism for CBT vs control
 - Naturally occurring groups
 - E.g., perfectionism in anorexia nervosa vs controls
- Associations among variables
 - Correlations/Regression Coefficients
 - E.g., correlation between perfectionism and depression

Answerable/Unanswerable Research Questions

▶ Unanswerable Research Questions

- What is the best strategy to reduce maladaptive perfectionism?
- How do we eliminate racism?

▶ Answerable Research Questions

- Are online interventions effective in reducing maladaptive perfectionism?
 - E.g., maladaptive perfectionism from pre-intervention to post-intervention
- Are males more racist than females?

Which Studies to Review?

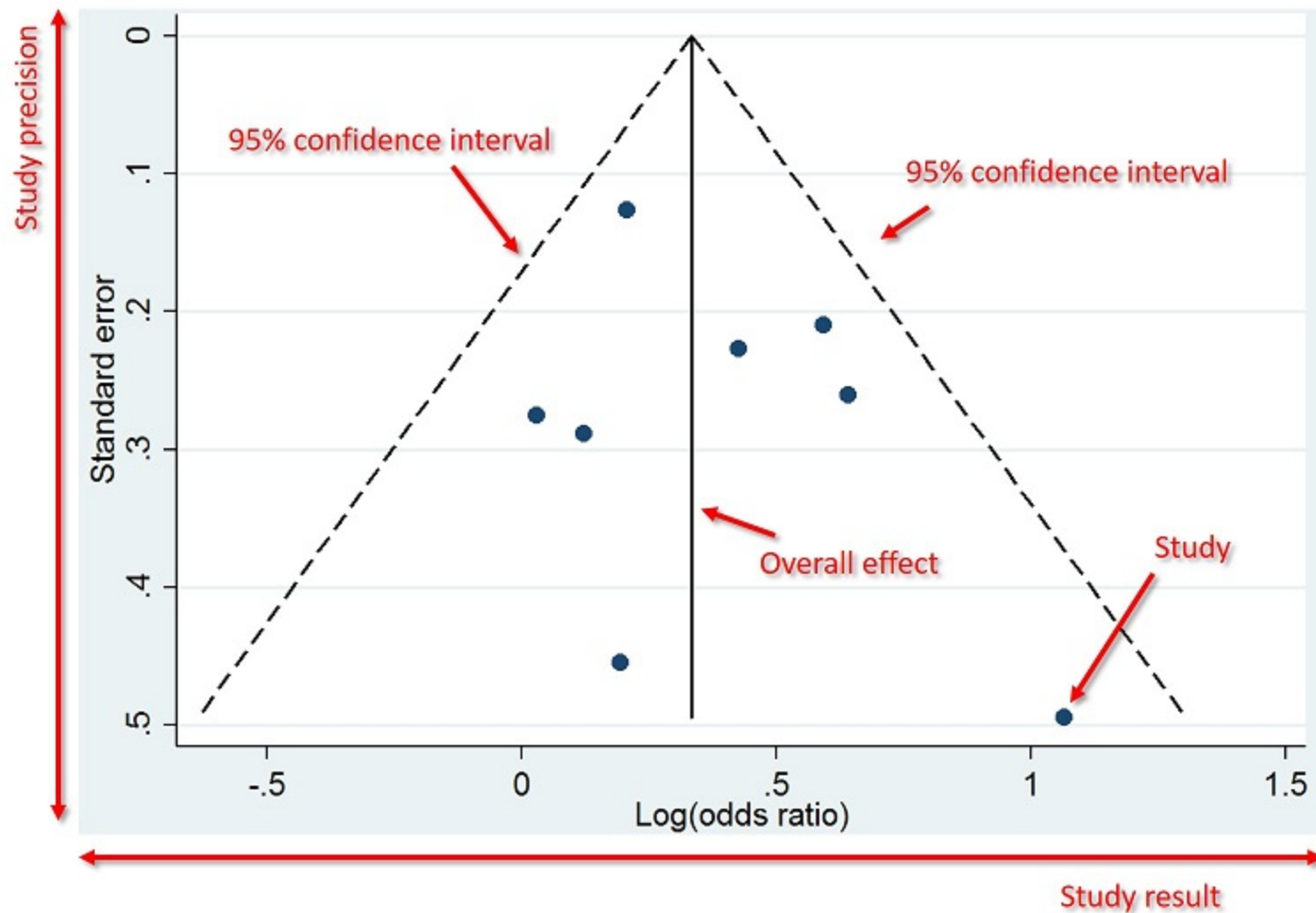
- ▶ Should be as inclusive as possible
 - Need to find ALL studies
 - Published studies are easy to find ... UNPUBLISHED STUDIES ARE NOT
 - The inclusion of unpublished studies helps to minimize the effects of *publication bias*
- ▶ Apples and Oranges
 - A priori inclusion and exclusion criteria must be laid out
 - It is imperative that the studies being meta-analyzed address the same research question

Exploring Publication Bias

▶ Funnel plot

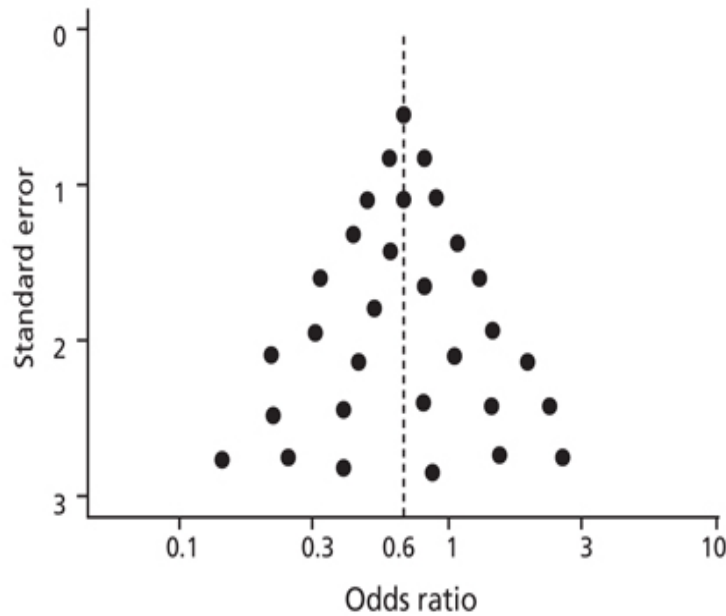
- A plot of the size of the effect of a study against the precision of a study
- Symmetrical funnel plots provide evidence of a lack of publication bias, where asymmetrical funnel plots highlight that publication bias might be present
 - E.g., if effects with low precision seem to all have larger effects then publication bias is likely

Funnel Plot

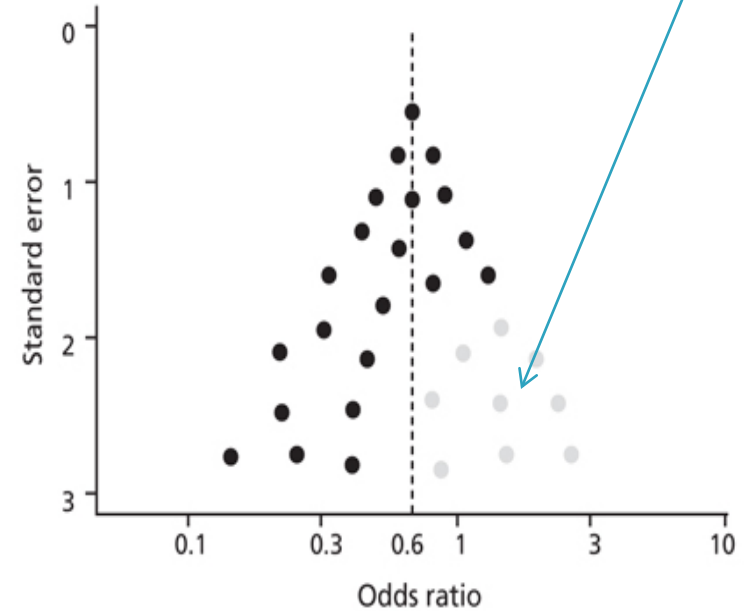


Symmetrical vs Asymmetrical Funnel Plot

No small N studies with OR between 1 and 3




A



B

Where To Find Studies

- ▶ Computerized bibliographic databases
 - Google Scholar, Psycinfo, Medline, ERIC
 - ▶ Authors working in the research domain
 - Personal websites (e.g., Researchgate, OSF, psyarchiv)
 - ▶ Conference programs
 - ▶ Dissertations
 - ▶ Reference lists from relevant articles
- 


What Information Should be Collected?

- ▶ Think about these long and hard before starting data collection ... it sucks to have to go back and recollect data
 - Publication details
 - Or specific location details for unpublished studies
 - Study design
 - Population details (N, characteristics)
 - Intervention/Design details
 - Operational Definitions of Variables
 - Demographics and other potential moderators
 - Outcomes
 - E.g., Means, SDs, correlations, regression coefficients, variability of coefficients, sample sizes

Why Assess the Validity of Studies?

- ▶ Lower quality studies can have biased outcome results
 - E.g., Allocation to Treatment/Control
 - Inadequate allocation concealment (e.g., investigators playing a role in allocation) exaggerated treatment effects by about 35% (Moher, 1998; Schulz, 1995)
 - E.g., Blinding
 - Lack of blinding of subjects exaggerated treatment effects by 17% (Schulz, 1995), or increased the effect size by about a half a SD (Hróbjartsson et al., 2014)

Where Can Bias be Introduced into Studies?

- Selection bias
 - Allocation bias
 - Confounds
 - Blinding
 - Data collection methods
 - Withdrawals and drop-outs
 - Statistical analysis
 - Intervention integrity
- ▶ Summary: Lots of ways that bias can be introduced into research
- 

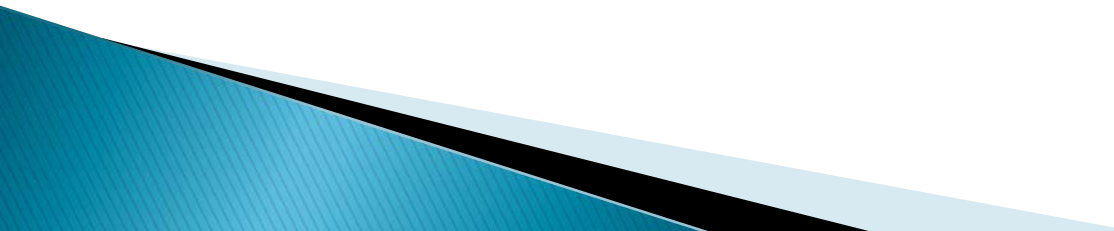
Assessing the Validity of a Study

- ▶ The most common way to assess and report study quality has been using a composite, numerical scoring instrument
 - Many different quality assessment instruments are available, with most designed for randomized clinical trials
- ▶ E.g., Jadad Score for Experiments (0–3)
 - Was the study described as randomized?
 - Was the study described as double-blind?
 - Was there a description of withdrawals and dropouts?

Methodological Quality Dilemma

- Include or exclude low quality studies?
 - The findings of all studies are potentially in error (methodological quality is a continuum, not a dichotomy)
 - Being too restrictive may limit ability to generalize
 - Being too inclusive may weaken the confidence that can be placed in the findings
 - Methodological quality is often subjective
 - You must strike a balance that is appropriate to your research question
- When including low quality studies you can weight effects by study quality or explore study quality as a moderator

Level of Replication

- ▶ Replications can range from “conceptual” replications to “pure” or “direct” replications
 - Direct replications are the repetition of an experimental procedure to as exact a degree as possible, whereas a conceptual replication is the use of different methods/procedures to repeat the test of a hypothesis
 - ▶ You must be able to argue that the collection of studies you are meta-analyzing examine the same relationship
 - ▶ The closer to pure replications your collection of studies, the easier it is to argue comparability of the effect from each study
- 

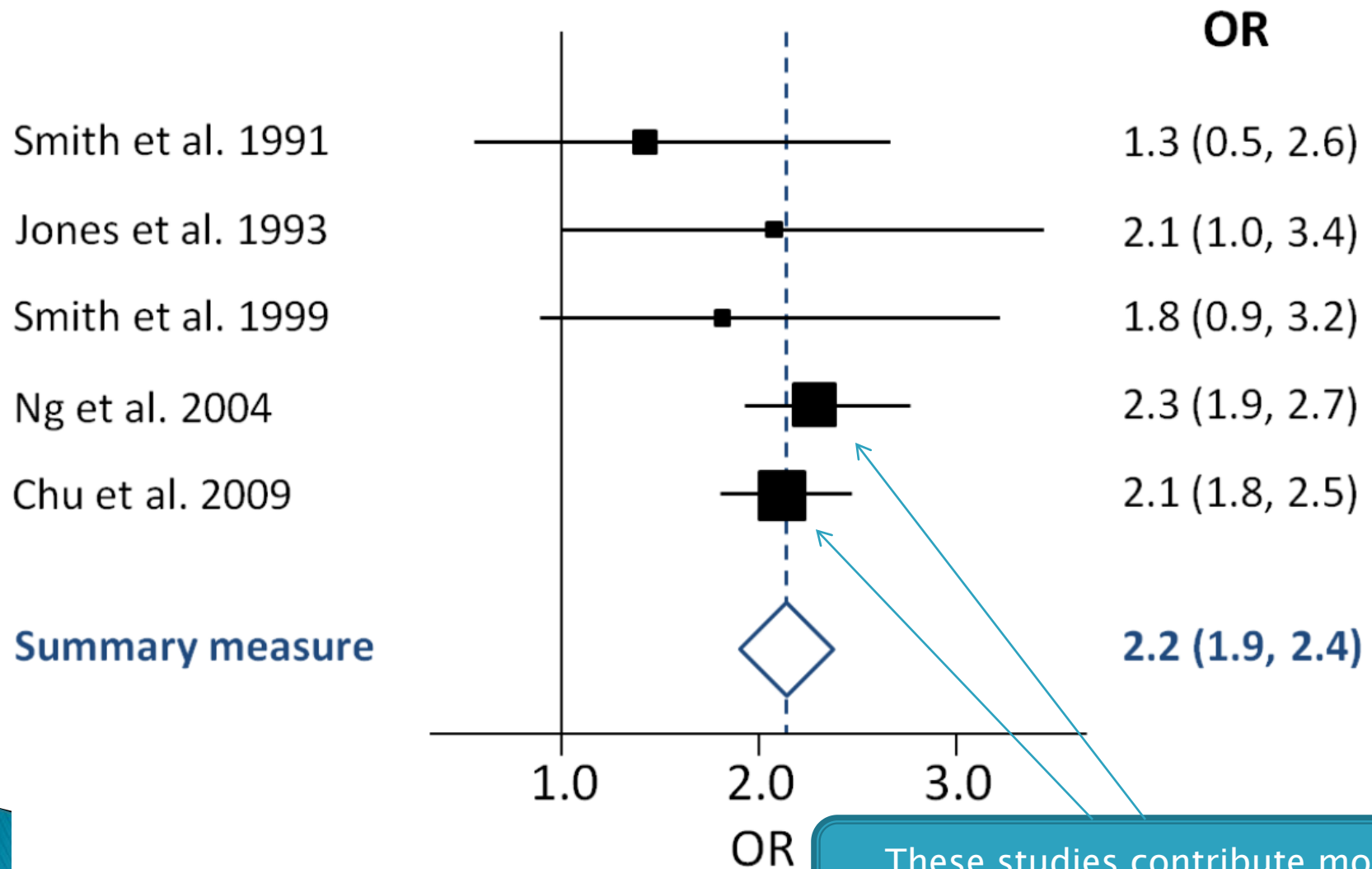
Effect Size in Meta-Analysis

- ▶ Effect size is the “dependent variable”
 - Standardizes findings across studies such that they can be directly combined/compared
 - A standardized index must be comparable across studies, represent the magnitude and direction of the relationship of interest, and be independent of sample size
 - e.g., standardized mean difference, correlation, odds-ratio
 - It is also possible to use unstandardized effect sizes, but this requires that the exact same variables are used in each study (and that no transformations, modifications, etc. were made to any variables)

Forest Plot

- ▶ A visual representation of the effect sizes (and confidence intervals for the effect sizes) of the multiple studies included in a meta-analysis
 - All effects must be measured in the same metric, e.g., correlation
 - It is often straightforward to transform from one effect size to another
- ▶ The area of the effect size icons (usually squares) indicates the “weight” of the study to the combined effect
 - E.g., larger N studies have a higher weight
- ▶ The plot also shows the effect size (and confidence interval for the effect size) of the combined effect across studies

Forest Plot Example – Odds Ratios



These studies contribute more information to the combined effect

Fixed Effects vs Random Effects

- ▶ There are two popular models available for conducting a meta-analysis
 - In other words, two models available for arriving at a “combined” measure of effect size
 - Fixed Effects Model
 - Assumes that all the studies investigated the same population, and therefore estimate the same population effect size
 - Highly questionable
 - Random Effects Model
 - Allows for the possibility that the studies investigated somewhat different populations, and therefore estimate different population effect sizes

Fixed Effects vs Random Effects

- ▶ It is difficult to imagine a setting in which multiple studies conducted in different locations, with different samples, and with potentially different measures are all studying the same population (and thus after a single population effect size)
- ▶ The random effects model is more realistic and provides a basis for understanding the heterogeneity of effect sizes
 - Further, the models give the same answer if there is only a single population, so it is hard to find a reason for a researcher to prefer a fixed effects model

Fixed Effects Meta-Analysis

- ▶ For a set of S effect size measures (γ)

- $\hat{\gamma}_F = \frac{\sum_{i=1}^S w_i \hat{\gamma}_i}{\sum_{i=1}^S w_i}$

- $w_i = \frac{1}{s^2(\hat{\gamma}_i)}$

- $s^2(\hat{\gamma}_F) = \frac{1}{\sum_{i=1}^S w_i}$

This info is used to generate a mean effect size and a CI around the mean effect size

Random Effects Meta-Analysis

- ▶ For a set of S effect size measures (γ)

- $\hat{\gamma}_R = \frac{\sum_{i=1}^S w_i \hat{\gamma}_i}{\sum_{i=1}^S w_i}$

- $w_i^* = \frac{1}{s^2(\hat{\gamma}_i) + \tau^2}$

- $\tau^2 = \frac{Q - (S-1)}{\sum_{i=1}^S w_i - \frac{\sum_{i=1}^S w_i^2}{\sum_{i=1}^S w_i}}$ for $Q > S-1$

- $Q = \sum_{i=1}^S w_i (\hat{\gamma}_i - \hat{\gamma}_F)^2$

- $s^2(\hat{\gamma}_R) = \frac{1}{\sum_{i=1}^S w_i^*}$

Studies are
weighted lower
when their effect

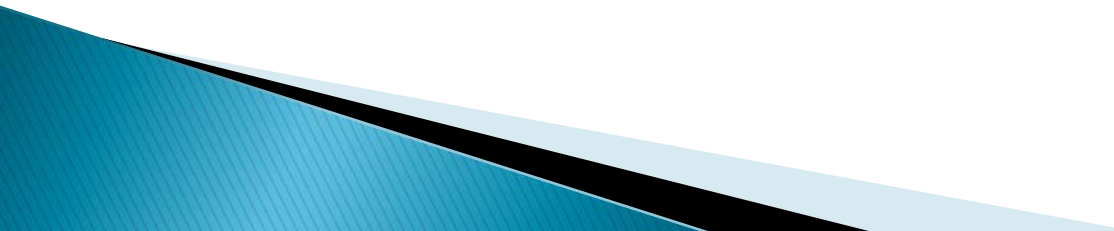
Heterogeneity of Effect Sizes

- ▶ A simple goodness-of-fit test can be used to test for excessive heterogeneity
 - $Q \sim \chi^2_{df=S-1}$
 - We reject the null that there is no population heterogeneity if $Q \geq \chi^2_{\alpha, df=S-1}$
- ▶ The problem with this approach is that the test has low-power when S is small

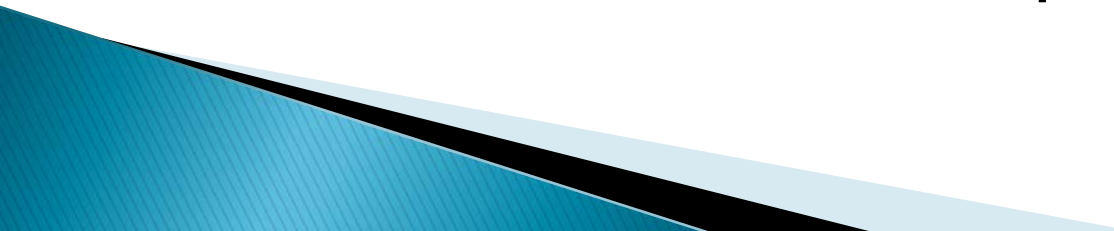
Heterogeneity of Effect Sizes

- ▶ A better approach to quantifying heterogeneity is to use an effect size measure
- ▶ $I^2 = \frac{Q-S+1}{Q}$
- ▶ I^2 ranges from 0 to 1, with larger values indicating more heterogeneity

Summary: Steps of a Systematic Review/Meta-Analysis

- ▶ Specify your research question/effect of interest
 - ▶ Find studies that investigate the effect of interest using inclusion/exclusion criteria
 - ▶ Extract all necessary information from the studies
 - ▶ Assess the validity of the studies
 - ▶ Assess risk of publication bias
 - ▶ Estimate the weighted combined effect size and CI for the effect size
 - ▶ Explore moderators of the variability in effect sizes
 - ▶ Interpret the findings
- 

Strengths of Meta-Analysis

- ▶ Imposes strict procedures on the process of summing up research findings
 - ▶ Represents findings in a more sophisticated manner than conventional reviews
 - ▶ Capable of finding relationships across studies that are obscured in other approaches or without amalgamation
 - ▶ Capable of detecting moderators of effects
 - ▶ Can handle a large numbers of studies, which would be difficult in a qualitative review
- 

Weaknesses of Meta-Analysis

- ▶ Requires a lot of effort!
- ▶ Mechanical aspects don't lend themselves to capturing more qualitative distinctions between studies
- ▶ “Apples and oranges”
 - Comparability of studies is often in the “eye of the beholder”
- ▶ Most meta-analyses include “blemished” studies
- ▶ Selection bias possesses continual threat
 - E.g., Null finding studies are hard to find

Conclusion: Why Meta-Analysis?

- ▶ Focuses on effect sizes, not statistical significance
- ▶ Combines multiple studies for a more precise estimate of the effect size
- ▶ Provides a rationale for small-N research
 - I.e., the results will be combined with other studies for a more precise estimate of the effect size

Example Meta-Analysis

DOI: 10.1002/eat.23009

International Journal of
EATING DISORDERS

REVIEW

Anorexia nervosa and perfectionism: A meta-analysis

Sophie C. Norris  | David H. Gleaves | Amanda D. Hutchinson

How to cite this article: Norris SC, Gleaves DH, Hutchinson AD. Anorexia nervosa and perfectionism: A meta-analysis. *Int J Eat Disord*. 2019;1-11. <https://doi.org/10.1002/eat.23009>

Step 1: Specify Research Question

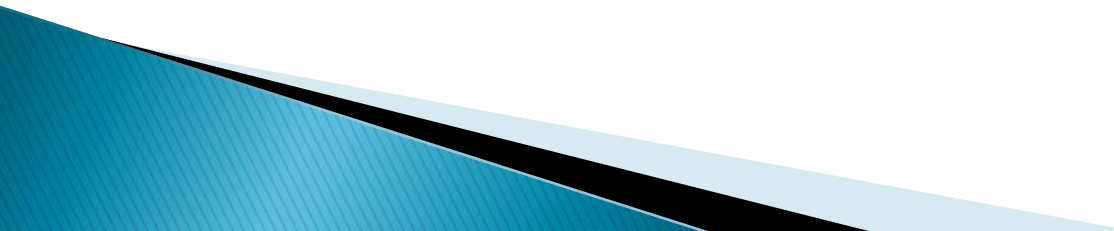
The aim of this study was to synthesize published research comparing perfectionism scores in those diagnosed with AN, with perfectionism scores of a non-clinical comparison group, a non-AN ED group, and PC group. Effect sizes were calculated, representing the

For the presentation I will
just focus on AN vs Non-
clinical Comparison

Step 2: Locate Studies that meet Inclusion/Exclusion Criteria

2.1 | Method design

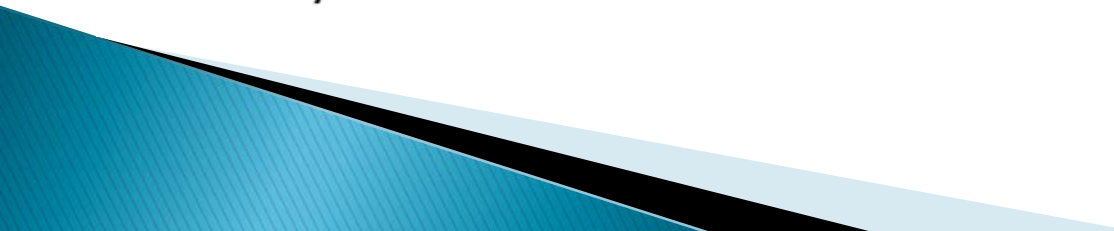
The research was conducted according to the PRISMA guidelines (Liberati et al., 2009), and we identified research papers that compared perfectionism scores in those diagnosed with AN and either a non-clinical comparison group, people diagnosed with a non-AN ED, or people diagnosed with another psychiatric disorder (i.e., other *DSM* diagnoses). The search identified relevant studies that met the following inclusion criteria.



Step 2: Locate Studies that meet Inclusion/Exclusion Criteria

2.1.1 | Inclusion criteria

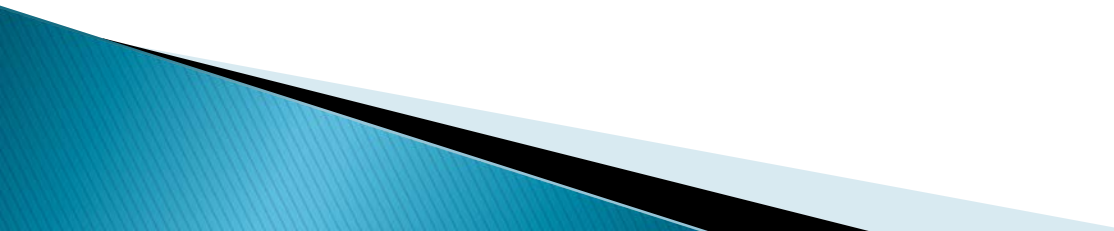
Studies that (a) included participants who were diagnosed with AN; and with either non-clinical comparison group, people diagnosed with a non-AN ED, or people diagnosed with another diagnosed psychiatric disorder, in accordance with the *DSM III, IV, or 5* criteria; (b) were peer-reviewed articles; (c) were empirical works; (d) were published in English; and (e) provided relevant statistics for perfectionism scores to allow calculation of effect size (e.g., *M*, *SD*, or *t*-test), were included in the analysis.



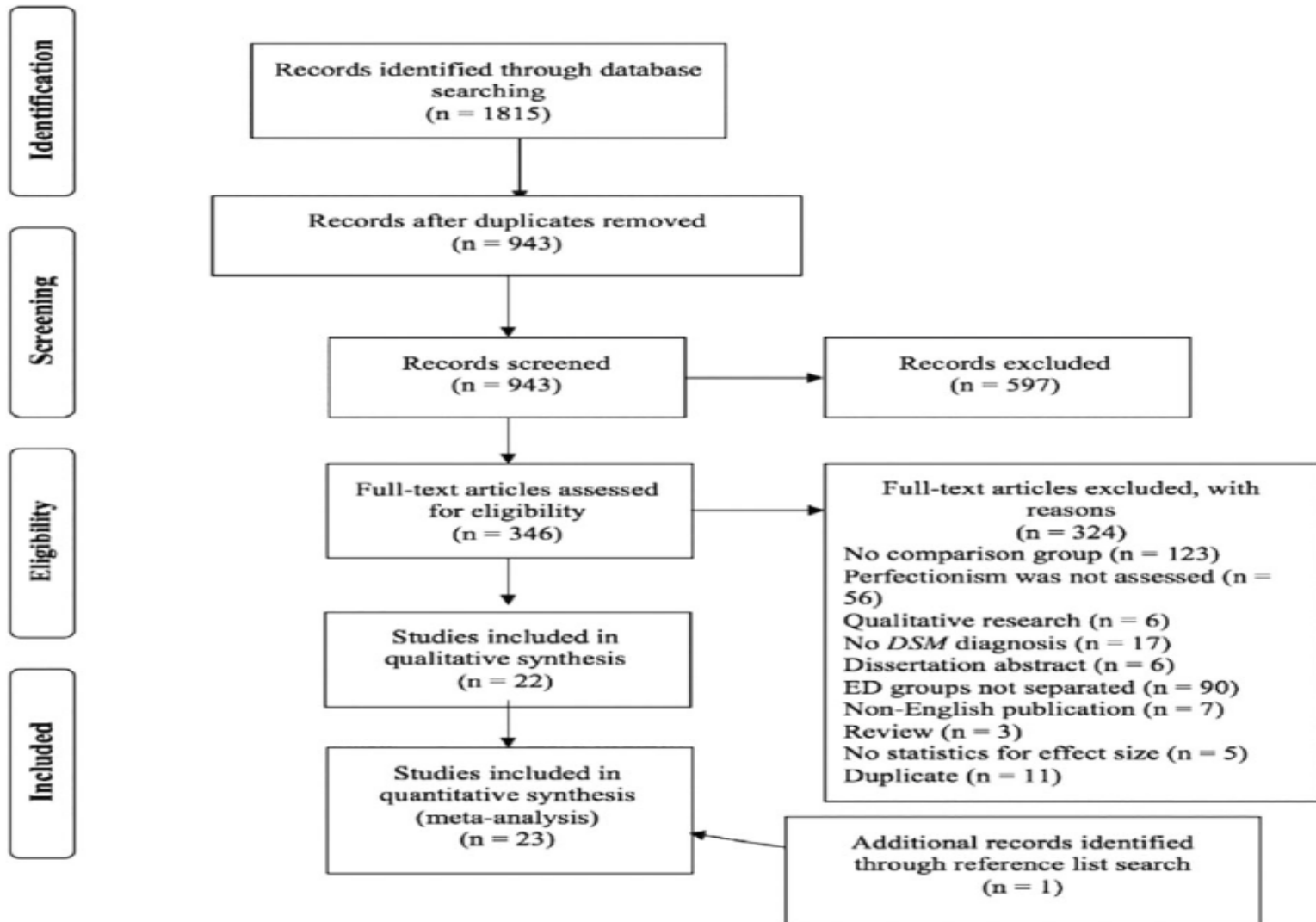
Step 2: Locate Studies that meet Inclusion/Exclusion Criteria

2.1.2 | Exclusion criteria

Studies that (a) had no clear diagnosis of AN, a non-AN ED, or another psychiatric disorder in accordance with the *DSM* criteria; (b) were not published in English; (c) provided no comparison group; (d) were meta-analyses or systematic reviews; (e) were case studies; or (f) either included insufficient results reported for calculation of effect size or results not available from authors, were not included.



Step 2: Locate Studies that meet Inclusion/Exclusion Criteria



Step 3: Extract Study Information

2.2.2 | Data extraction

The data extracted from each study were, where applicable, year published, country, age of participants, gender, number of participants in each group, version of *DSM* used for diagnoses, *DSM* diagnosis, specifically how the *DSM* diagnosis was reached, measure used for assessing perfectionism, and any group statistics reported used to calculate the effect size.

Step 3: Study Info Results

Study	Country	Race/ethnicity	Perfectionism measure	AP, MP, or both captured	Methodological quality score	DSM version	Method of diagnosis	AN group (n)	Comparison group (n)	N
Bachner-Melman et al. (2006)	IL	Not reported	CAPS	Both	22/22	IV	SCID	AN (31)	Non-clinical (248)	279
Bachner-Melman et al. (2007)	IL	Not reported	CAPS	Both	18/22	IV	SCID	AN (17)	Non-clinical (242)	259
Casper et al. (1992)	US	Not reported	EDI-P	MP	18/22	III-R	SCID	AN-BP (19) AN-R (12)	BN (19) Non-clinical (19)	50
Castro-Fornieles et al. (2007)	ES	Not reported	CAPS	Both	22/22	IV	Clinical interview	AN (75)	BN (33) PC (86) Non-clinical (213)	407
Dalle Grave, Calugi, and Marchesini (2008)	IT	Not reported	EDI-P	MP	19/22	IV	SCID	AN-BP (30) AN-R (35)	BN (28)	93
Davis and Scott-Robertson (2000)	US	Not reported	MPS	AP	18/22	IV	Not reported	AN (46)	Non-clinical (22)	68
Davies, Liao, Campbell, and Tchanturia (2009)	UK	Not reported	F-MPS	MP	19/22	IV	SCID	AN (30)	BN (26) Non-clinical (51)	107
Fassino, Amianto, and Abbate-Daga (2009)	IT	Not reported	EDI-P	MP	22/22	III	SCID	AN-BP (30) AN-R (38)	BN (35) Non-clinical (54)	159
Fassino, Piero, Gramaglia, and Abbate-Daga (2004)	IT	Not reported	EDI-P	MP	22/22	IV	SCID	AN-BP (61) AN-R (61)	BN (104)	226
Halmi et al. (2000)	US	Not reported	F-MPS	MP	20/22	IV	Not reported	AN-BP (60) AN-R (146)	Non-clinical (44)	250
Jiménez-Murcia et al. (2007)	ES	Spanish	EDI-P	MP	20/22	IV	SCID	AN (30)	BN (30) PC (30)	90
Kim et al. (2010)	KR	Korean & British	CRF-Q	MP	18/22	IV	Semi-structured interview	AN (52)	Non-clinical (108)	202
Moor, Vartanian, Touyz, and Beumont (2004)	AU	Not reported	EDI-P	MP	19/22	IV	Not reported	AN (27)	BN (23) Non-clinical (25)	75

Subset of studies ...

Step 4: Study Validity

2.3.1 | Methodological quality

We addressed the risk of bias based on methodological quality using the Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of Fields checklist (Kmet, Lee, & Cook, 2004). The checklist covers criteria such as study design, outcome measures, sample size, and if the results and conclusion are congruent. Responses can be “yes”, “partial”, or “no” and these responses are worth either two, one, or zero points, respectively. With 11 items on the checklist used for quantitative research, each study assessed could score a maximum of 22 points, indicating strong methodological quality.

Step 4: Study Validity Results

TABLE 1 Studies included in the meta-analysis of perfectionism levels of AN and comparison groups

Study	Country	Race/ethnicity	Perfectionism measure	AP, MP, or both captured	Methodological quality score
Bachner-Melman et al. (2006)	IL	Not reported	CAPS	Both	22/22
Bachner-Melman et al. (2007)	IL	Not reported	CAPS	Both	18/22
Casper et al. (1992)	US	Not reported	EDI-P	MP	18/22
Castro-Fornieles et al. (2007)	ES	Not reported	CAPS	Both	22/22
Dalle Grave, Calugi, and Marchesini (2008)	IT	Not reported	EDI-P	MP	19/22
Davis and Scott-Robertson (2000)	US	Not reported	MPS	AP	18/22
Davies, Liao, Campbell, and Tchanturia (2009)	UK	Not reported	F-MPS	MP	19/22
Fassino, Amianto, and Abbate-Daga (2009)	IT	Not reported	EDI-P	MP	22/22
Fassino, Piero, Gramaglia, and Abbate-Daga (2004)	IT	Not reported	EDI-P	MP	22/22
Halmi et al. (2000)	US	Not reported	F-MPS	MP	20/22
Jiménez-Murcia et al. (2007)	ES	Spanish	EDI-P	MP	20/22

Subset of studies ...

Step 5: Publication Bias

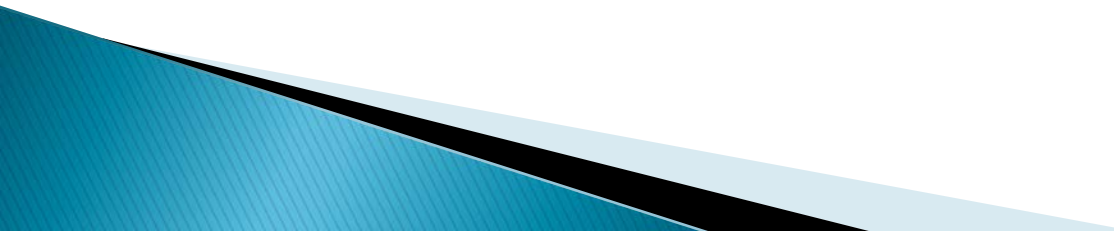
2.3.2 | Publication bias

The Fail-safe N is a statistical tool that addresses publication bias. The resulting calculation estimates the number of unpublished studies needed to make a statistically significant result no longer statistically significant (Rosenthal, 1979). We used the Fail-safe N to determine publication bias in the studies, as studies that produce a significant result are more likely to be published than non-significant results.

Forest plots and funnel plots were generated to visually inspect heterogeneity and publication bias in the results. The forest plot visually shows the heterogeneity, or differences in results, in the included studies. For a potential indicator of publication bias, a funnel plot

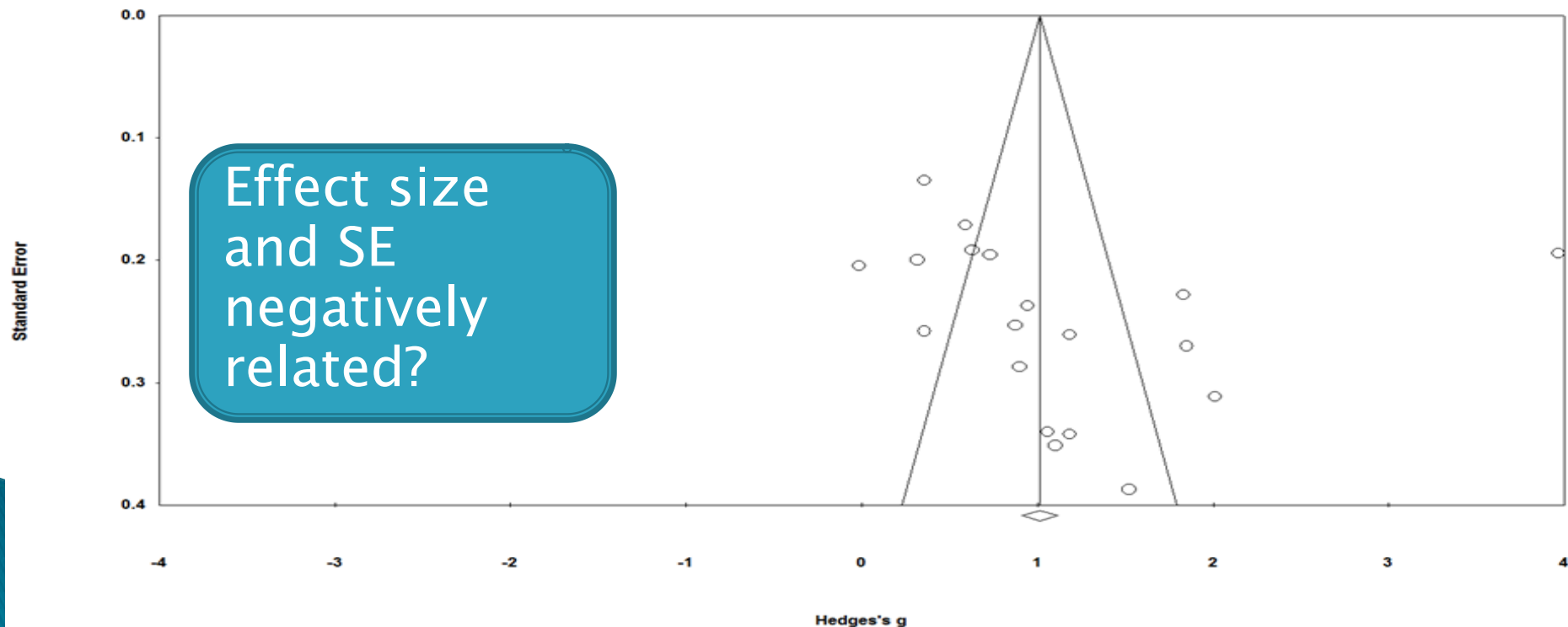
Step 5: Publication Bias Results

The Fail-safe N was only acceptable for the two non-clinical group comparisons, indicating it is unlikely there are enough unpublished studies with a statistically non-significant effect to make this result statistically non-significant. The Fail-safe N for the PC group was below the minimum required value, which suggests that it is possible that there are a number of studies in existence that could overturn the significance of this result.



Step 5: Publication Bias Results

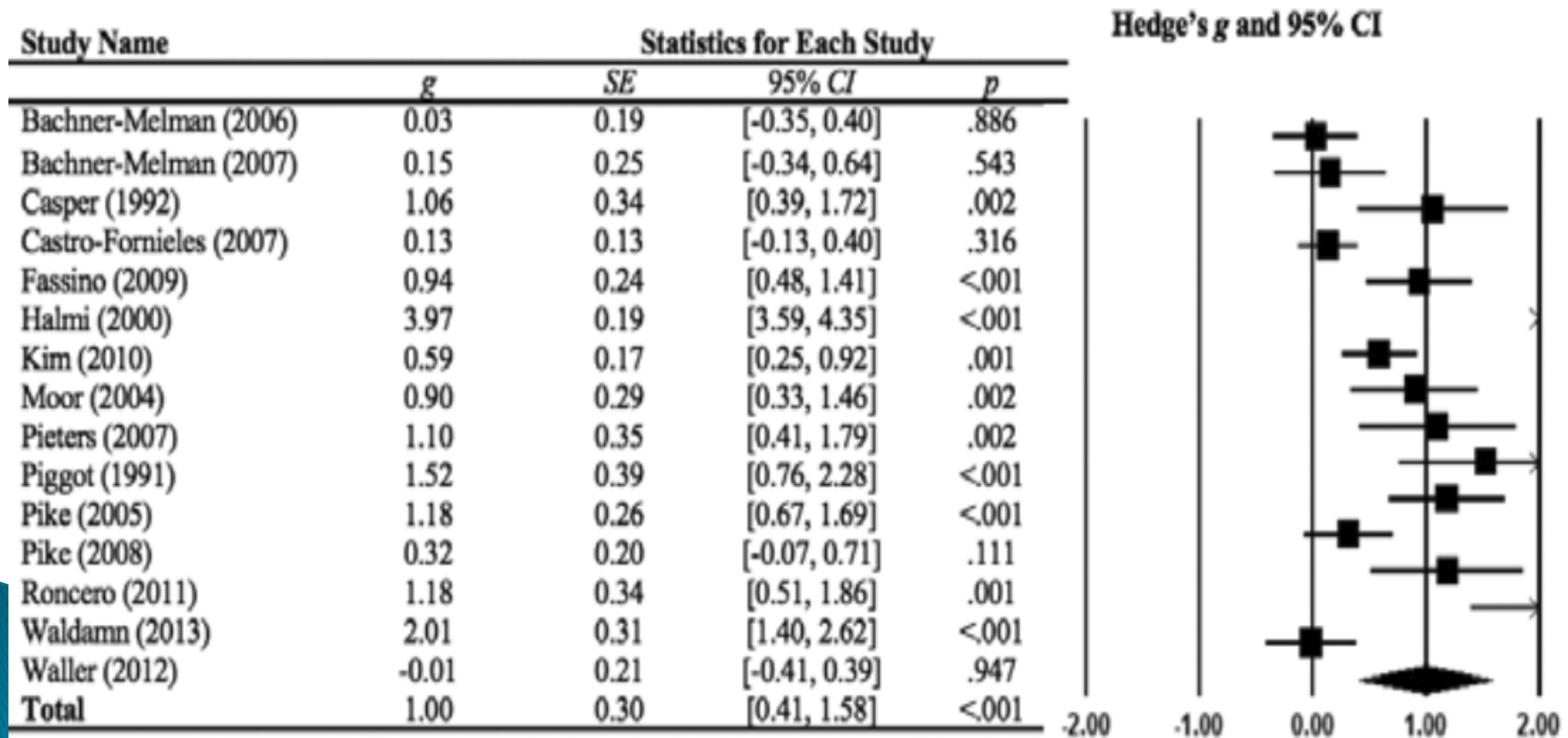
Funnel plots for the group comparisons were visually inspected and identified no asymmetry.



Step 6: Combine Effect Sizes

We conducted a random-effect analysis due to the assumed heterogeneity between the studies (there were varying types of perfectionism measures used, and the methodology of the studies varied; Borenstein et al., 2009).

AN vs
Control

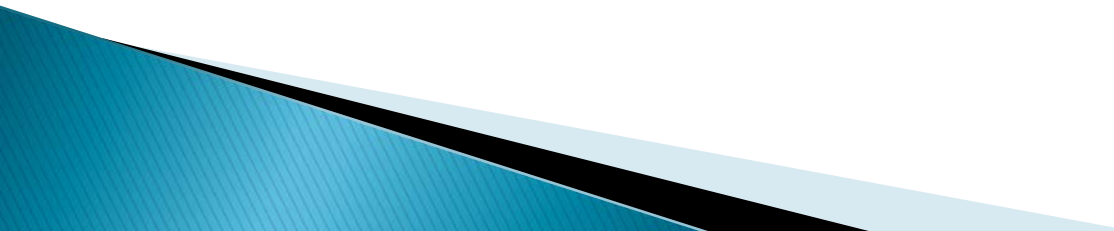


Step 7: Moderators

Furthermore, too few of the studies reported ethnicity data so we were unable to include the variable as a moderator.

Step 8: Conclusions

There were no statistically significant differences in maladaptive perfectionism between individuals diagnosed with AN and BN. The results from the meta-analysis also supported the hypothesis that the AN group was more perfectionistic compared to the non-clinical group, and the effect size was large. This result was the same for both maladaptive and adaptive perfectionism.



General Conclusions

- ▶ Meta-analysis is a valuable tool for combining results (effect sizes) from multiple studies and providing a sense of the overall magnitude of the effect
 - ▶ Researchers in Psychology are slowly warming up to the value of meta-analyses, and it is important that we are now familiar with meta-analyses in our fields
 - And conduct them when they are missing!
- 