HHS Public Access

Author manuscript

J Adolesc. Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

J Adolesc. 2019 December ; 77: 168–178. doi:10.1016/j.adolescence.2019.10.009.

Meta-analytic Approaches for Examining Complexity and Heterogeneity in Studies of Adolescent Development

Nicholas J. Parr[a,b], **Maria L. Schweer-Collins**[a,b], **Todd M. Darlington**[b,c], **Emily E. Tanner-Smith**[a,b]

[a]Department of Counseling Psychology and Human Services, University of Oregon. 5251 University of Oregon, Eugene, OR, 97403, United States of America.

[b]Prevention Science Institute, University of Oregon. 6217 University of Oregon, Eugene, OR, 97403, United States of America.

[c]Department of Psychology, University of Oregon. 1227 University of Oregon, Eugene, OR, 97403, United States of America.

Abstract

Introduction.—In the field of adolescent development, meta-analysis offers valuable tools for synthesizing and assessing cumulative research evidence on the effectiveness of programs, practices, and policies intended to promote healthy adolescent development. When examining the impact of a program implemented across multiple primary studies, variation is often observed in the methodological attributes of those primary studies, such as their implementation methods, program components, participant characteristics, outcome measurement, and the systems in which programs are deployed. Differences in methodological attributes of primary studies represented in a meta-analysis, referred to as complexity, can yield variation in true effects across primary studies, which is described as heterogeneity.

Methods.—We discuss heterogeneity as a parameter of interest in meta-analysis, introducing and demonstrating both graphical and statistical methods for evaluating the magnitude and impact of heterogeneity. We discuss approaches for presenting characteristics of heterogeneity in meta-analytic findings, and methods for identifying and statistically controlling for aspects of methodological complexity that may contribute to variation in effects across primary studies.

Results.—Topics and methods related to assessing and explaining heterogeneity were contextualized in the field of adolescent development using a sample of primary studies from a large meta-analysis examining the effectiveness of brief alcohol interventions for youth. We highlighted approaches currently underutilized in the field and provided R code for key methods to broaden their use.

Correspondence concerning this manuscript should be addressed to Nicholas J. Parr, Department of Counseling Psychology and Human Services, 5251 University of Oregon, Eugene, OR, 97403. Contact: nparr2@uoregon.edu.

Declaration of interest: none.

Conclusions.—By discussing various heterogeneity statistics, visualizations, and explanatory methods, this article provides the applied developmental researcher a foundational understanding of heterogeneity and complexity in meta-analysis.

Keywords

Meta-analysis; heterogeneity; methodological complexity; meta-regression; R

Introduction

Meta-analysis provides valuable tools for synthesizing and assessing cumulative research evidence (Lipsey & Wilson, 2001). In the field of adolescent development, meta-analyses have been used to investigate the effectiveness of programs, practices, and policies (hereafter described as programs) intended to promote healthy development over the course of adolescence (Card & Little, 2006; Clarke, 2006; Durlak, Weissberg, & Pachan, 2010; Horowitz & Garber, 2006; Weisz, McCarty, & Valeri, 2006). Some of these meta-analyses synthesize correlational or observational evidence of risk and protective factors for healthy adolescent development (e.g., Tanner-Smith, Wilson, & Lipsey, 2013), while others synthesize evidence from controlled evaluations of program effects to answer fundamental questions about whether a program is generally helpful, harmful, or ineffective (e.g., StrØm, Adolfsen, Fossum, Kaiser, & Martinussen, 2014; Tripodi, Bender, Litschge, & Vaughn, 2010). When assessing the impact of a program implemented in multiple primary studies drawn from a systematic review of available literature, observing differences in methodological characteristics across primary studies is common. Primary studies can differ in, for instance, implementation methods, program components, participant characteristics, outcome measurement, and the systems in which a program is deployed. Differences in methodological attributes of primary studies included in a meta-analysis, referred to here as complexity, can produce variation in true effects across primary studies, which is described as heterogeneity (Borenstein, Higgins, Hedges, & Rothstein, 2017; Higgins & Thompson, 2002). Heterogeneity can considerably complicate the interpretability of meta-analytic findings, and as a result, it is recommended practice to at minimum assess the magnitude of heterogeneity among the effects of primary studies included in a meta-analysis (Higgins, 2008; Higgins & Green, 2011; Moher et al., 1999). In addition to quantifying heterogeneity, statistical and graphical tools can be used to evaluate the impact of heterogeneity on an overall effect size estimate and to investigate whether specific sources of heterogeneity can be identified (Baker, White, Cappelleri, Kluger, & Coleman, 2009; Hardy & Thompson, 1998; Higgins & Thompson, 2002; Thompson, 1994; Viechtbauer, 2007). Utilizing these approaches to examine whether aspects of methodological complexity are sources of heterogeneity in meta-analytic findings can provide a more nuanced view of when, for whom, and under what conditions program effects may vary, and yield information that can be used to identify and address differences in outcomes and to tailor programs to optimize their effectiveness across settings (Lipsey & Wilson, 2001; Viechtbauer, 2007).

This article aims to provide the applied developmental researcher with a foundational understanding of heterogeneity in meta-analysis, including methods for quantifying and assessing the impact of heterogeneity, and for explaining heterogeneity resulting from

methodological complexity. Topics addressed include the historical and statistical background of heterogeneity as a parameter of interest in meta-analysis, graphical and statistical methods for evaluating the magnitude and impact of heterogeneity, and approaches for presenting characteristics of heterogeneity in meta-analytic findings.[1] We also discuss and demonstrate methods for identifying and statistically controlling for aspects of methodological complexity that may contribute to variation in effects across primary studies.[2]

Motivating Example

To place these topics in the context of adolescent development, key methods will be illustrated using a sample of primary studies from one of the largest meta-analyses to date of brief alcohol interventions (BAIs) for youth, which synthesized evidence from 190 randomized controlled trials and controlled quasi-experiments (Tanner-Smith & Lipsey, 2015; Tanner-Smith & Risser, 2016). BAIs are low-cost preventive interventions aimed at promoting change in alcohol use behaviors or their determinants, and while they are typically short-duration and administered in a single session, BAIs can vary in method of delivery, implementation setting, and in primary components. In the meta-analysis that serves as the motivating example for this article, participants in primary studies were adolescents and young adults ages 11–30, and included studies compared a BAI condition to an inactive comparison condition such as wait-list control, treatment-as-usual, or no treatment (Tanner-Smith & Lipsey, 2015). For the example analyses reported here, we synthesized post-intervention effects on alcohol-related consequences using the Rutgers Alcohol Problems Index (RAPI; White & Labouvie, 1989), and utilized the standardized mean difference between intervention and comparison conditions as the effect size measure. In the sections that follow, approaches for examining and explaining heterogeneity are demonstrated using the *metafor* package for R (version 2.0-0; Viechtbauer, 2010, 2017).[3]

Understanding Heterogeneity

The principal aim of many meta-analyses is to summarize the findings of multiple primary studies using an overall or average effect estimate. In the context of evaluating the effectiveness of a program using meta-analysis, the program will have an observed effect in each primary study that is assumed to reflect the true (unobserved) effect in the population; however, the true effect of the program may vary considerably across contexts, suggesting that there may be multiple true effects in the population. In the scenario where true effects

[1]In this article we confine our focus to meta-analysis of aggregate data, historically the most widely used form of meta-analysis. We direct the reader interested in meta-analysis of individual participant data to more comprehensive texts on that topic, including Cooper and Patall (2009), Riley, Lambert, and Abo-Zaid (2010), and Stewart, Tierney, and Clarke (2011).

[2]Although the data example used in this article synthesizes standardized mean difference effect sizes from primary studies using between-group designs to assess intervention effects, all of the statistical and graphical tools discussed in this article can be widely applied to meta-analyses, regardless of the type of research design used in the primary studies, and regardless of the effect size metric employed. For instance, meta-analyses may focus on synthesizing effects from studies that use single-case designs (Burns, 2012; Shadish, Hedges, & Pustejovsky, 2014; Valentine, Tanner-Smith, Pustejovsky, & Lau, 2016), regression discontinuity designs (Valentine, Konstantopoulos, & Goldrick-Rab, 2017), designs assessing diagnostic test accuracy (Deeks, Bossuyt, & Gatsonis, 2013), or a range of other designs (Cooper, Hedges, & Valentine, 2019). We direct readers to the aforementioned resources for in-depth discussion of unique considerations associated with synthesizing evidence from these types of designs.

[3]Complete R code is provided in the supplementary materials to this article, found at osf.io/q3nj5. Readers interested in the application and interpretation of illustrated procedures in *metafor* may find it informative to view the supplement alongside this article. R version 3.5.1 (R Core Team, 2018; RStudio Team, 2018) was used for all illustrated procedures.

vary insubstantially across contexts (that is, the true effects are homogeneous), an average effect size estimate can be a useful representation of the common true effect in the population. Conversely, in the situation where true effects vary considerably between studies (i.e., there is not a common true effect), the average effect size estimate does not have the same utility: true effects could be dispersed broadly around the average effect, ranging in magnitude (from much smaller to much larger), direction (from helpful to harmful), or both. This variation in true effects is heterogeneity, and it has clear implications for the interpretability of the overall or average effect reported in a meta-analysis. Assessing these implications, however, is complicated by the reality that we cannot directly synthesize the true effect in each primary study because each true effect is measured with error. Consequently, a meta-analysis summarizes *observed* effects – true effects accompanied by measurement error present in each study – rather than true effects alone.

Reliance on observed effects at the primary study level necessitates that true effects, as well as any variation among those true effects (heterogeneity), must be estimated. Over the last two decades, estimation tools for heterogeneity have become broadly available, and at the same time, there has been growing acceptance that heterogeneity in meta-analysis is often inevitable and should be anticipated (Higgins, 2008; Lorenc et al., 2016). These developments have accompanied the widespread use of random- and mixed-effects statistical models for meta-analysis, which require estimation of between-study variability in true effects, in contrast to a fixed-effect approach that assumes all primary studies share a common, or homogeneous, true effect (Riley, Higgins, & Deeks, 2011).[4] In the field of adolescent development in particular, it is typically unreasonable to assume that all primary studies in a meta-analysis are drawn from a single homogeneous population (the assumption underlying fixed-effect models). For this assumption to hold, all included studies would need to share virtually identical methodological, contextual, and participant characteristics, and there could be no variation in the true effect of a program or intervention beyond that expected by chance.[5] Because a more realistic assumption is often that heterogeneity is present to some degree in a sample of primary studies, random-effects models for meta-analysis are now generally recommended (Borenstein, Hedges, Higgins, & Rothstein, 2010; Lipsey & Wilson, 2001; Tanner-Smith & Grant, 2018). Alongside broader use of random-effects models, various approaches to expressing heterogeneity and assessing its impact on an overall effect size estimate have been developed, and it is now common for meta-analysts to examine heterogeneity in an effort to understand what factors may explain variation in effects across studies, contexts, and participants. There has also been increased recognition that when these factors are malleable (e.g., aspects of intervention implementation such as duration or delivery method), insight garnered from investigating heterogeneity in effects can be especially valuable to future research and practice.

[4]Random-effects models also typically provide more generalizable findings than fixed-effect models, because they explicitly model between-study variability, and consequently, the model estimates can be assumed to more accurately reflect a broader population of primary studies and their effects than those included in a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010).

[5]Fixed-effect models may be plausible in some situations, however, such as when researchers have conducted several nearly identical studies and are seeking to assess the common effect of a given treatment for a small, relatively homogeneous set of samples. Fixed-effects (plural) models may be used to examine between-study variation among a sample of primary studies when there is not an interest in generalizing findings to a broader population of studies. Hedges and Vevea (1998) and Borenstein et al. (2010) discuss the strengths, limitations, and applications of random-effects and fixed-effect(s) models for meta-analysis in greater detail.

Quantifying and Assessing Heterogeneity

Before investigating sources of heterogeneity in an overall effect size estimate, an important initial step is estimating the amount or magnitude of between-study variation (heterogeneity) across the true effect sizes in the population. The amount of heterogeneity is typically expressed as the parameter τ^2 (tau-squared), and its square-root, τ, represents the standard deviation of true effects between studies. Estimation of τ^2 can be accomplished using method-of-moments estimation or iterative approaches such as restricted maximum-likelihood estimation.[6] Once estimated, τ^2 and τ take the unit of the effect size measure employed in a meta-analysis. This attribute can be useful in some situations, while in others, it can complicate the interpretation of the statistics as standalone measures of heterogeneity. For instance, when an effect size measure is a standardized mean difference (e.g., Cohen's *d*), the estimate of τ is in the same units as the standardized mean difference, and as a result provides a straightforward index of the dispersion (standard deviation) of true effects around the overall effect size estimate. In other cases, such as when the effect size measure is a log odds or risk ratio, the estimate of τ may not offer as intuitive an interpretation because it is in the form of log-units (Borenstein et al., 2017).[7] In a later section, we present an approach for expressing the dispersion of true effects around an overall effect size estimate that utilizes τ^2, but is directly interpretable regardless of the effect size measure used in a meta-analysis.

Common Presentations of Heterogeneity

In addition to τ^2, which represents the amount of heterogeneity in true effects, there are numerous statistical and graphical approaches that can be used for presenting characteristics of heterogeneity in meta-analytic findings. Those we introduce below are some of the most commonly used approaches. Our aim is to equip the reader with a working knowledge of the meaning, intended use, and limitations of these tools, so that when encountered, they may be interpreted accurately. We also discuss and illustrate the use of prediction intervals for expressing heterogeneity, an approach that offers great utility but is currently underutilized in meta-analyses in the field of adolescent development.

Cochran's *Q*, often referred to as simply *Q*, summarizes the variation between each primary study's estimated effect and the overall effect size estimate. In a random-effects meta-analysis, *Q* is weighted by each study's precision (the inverse of its estimated error variance) and by between-study variation (τ^2). Given these properties, *Q* has frequently been used to test whether the amount of between-study variation (heterogeneity) is significantly different from zero (with estimates of zero suggesting homogeneity, or no variation in true effect between studies in excess of what would be expected due to chance).[8] Although a significance test that can be used to test for the presence or absence of heterogeneity is

[6]Importantly, estimation of τ^2 can be biased by the number of primary studies and other sample characteristics depending on the method used to calculate the statistic. In simulation studies, the restricted maximum-likelihood (REML) estimator has been found to be less susceptible to underestimating heterogeneity and other biases than both full maximum-likelihood estimation and moment-based estimators (Langan et al., 2018; Viechtbauer, 2005). Comparative strengths and drawbacks of different estimation approaches for τ^2 are discussed in Langan, Higgins, and Simmonds (2017) and Langan et al. (2018).

[7]Related to these considerations, estimates of τ^2 cannot be easily compared across meta-analyses that use different effect size measures (Higgins & Thompson, 2002).

conceptually appealing, a serious limitation of the Q test is that it is underpowered to detect departures from homogeneity when the total number of included primary studies is small, and overpowered to detect heterogeneity when the number of studies is large (Hardy & Thompson, 1998; Higgins & Thompson, 2002; Mittlböck & Heinzl, 2006). This can result in clinically important heterogeneity that remains undetected, or clinically unimportant heterogeneity that is detected. In addition to this statistical significance test, another common use of the Q statistic is in a visual presentation of effect size heterogeneity known as a Baujat plot (Baujat, Mahé, Pignon, & Hill, 2002). Figure 1 illustrates a Baujat plot using the example BAI meta-analysis data. The primary studies that have the greatest variation from the overall effect size estimate and the most substantial contribution to the estimate are located in the upper-right corner of the plot (Anzures-Cabrera & Higgins, 2010). While a Baujat plot may be useful for identifying specific primary studies whose effects vary most substantially from those in other included studies, inspection of the plot may lead to the erroneous conclusion that the most severely-outlying studies should be removed to reduce heterogeneity. Such a practice is discouraged because doing so amounts to manipulating the eligibility criteria of a meta-analysis (Higgins, 2008). Below we discuss methods that allow for sources of heterogeneity to be identified and taken into account without study removal, such as meta-regression.

The heterogeneity statistic I^2 addresses some of the limitations of Q as a measure of heterogeneity, in particular that Q is influenced by the number of primary studies included in the meta-analysis (Higgins & Thompson, 2002; Mittlböck & Heinzl, 2006). The I^2 statistic is a relative heterogeneity indicator, in that it expresses the proportion of true between-study variation relative to total observed variation (which is comprised of true between-study variation and within-study error). Because it is a proportion, I^2 takes the form of a fixed interval of 0–1 or 0–100%. A value of 0 indicates that all observed variation is attributable to within-study variation (i.e., due to error, reflecting homogeneity between studies), whereas an I^2 value of 1 (or 100%) denotes that all observed variation is due to true between-study variation (heterogeneity). The I^2 statistic has come into widespread use but has sometimes been presented as a measure of the *magnitude* of heterogeneity and been used to categorize heterogeneity into low, moderate, or high levels. These uses have led to some confusion in the statistic's interpretation (Borenstein, 2018; Borenstein et al., 2017; Higgins, 2008). Namely, I^2 conveys the relative *composition* of variation in an overall effect size estimate, and not the *extent* (i.e., range) of variation in effect sizes between primary studies. As a result, the statistic is not suited for use as an overall statement of the impact of heterogeneity because it does not indicate the absolute magnitude of heterogeneity, and by extension, does not indicate whether effect sizes vary minimally or considerably between studies (Borenstein, 2018). For example, an I^2 value of 80% indicates that approximately 80% of the observed variation in an effect size estimate is due to true between-study variation (heterogeneity), and yet with no other information we cannot know if that value reflects one of two situations: the case where between-study variation makes up 80% of a *substantial*

[8]When homogeneity is present, the Q-test asymptotically follows a χ^2-distribution, with degrees of freedom equal to $k-1$, where k is the number of primary studies (Hedges & Vevea, 1998). As a result, a significant Q test is interpreted as a significant presence of variation over and above within-study variation (error). Importantly, however, the assumption of a null χ^2-distribution of Q is reasonable only when sample sizes of primary studies are large (Hedges, 1982; Hoaglin, 2015; Kulinskaya, Dollinger, & BjØrkestØl, 2011).

amount of total variation, or the case where between-study variation makes up 80% of an *inconsequential* amount of total variation. In both of these cases, between-study and within-study variation can contribute to the effect size estimate in the same proportion (to one another), but with the two very different conclusions of considerable heterogeneity in the former case, and minimal heterogeneity in the latter case.

To more concretely illustrate these points, including limitations of the common heterogeneity indicators we have overviewed, we can turn to the example BAI meta-analysis data. For this example, a random-effects model was used and τ^2 was estimated using the restricted maximum-likelihood approach (the default in the *metafor* package; see Footnote 6).We used an $\alpha = .05$ to assess statistical significance. According to the model (see Supplement for full output), the overall effect size estimate of the post-intervention standardized mean difference (Cohen's *d*) in alcohol-related problems between intervention and comparison conditions is −0.21, 95% CI [−0.28, −0.13]. This estimate indicates that, in general, BAIs are associated with lower levels of alcohol-related problems, relative to control conditions. For this estimate, just over one-third of the observed variation in the estimate is due to true between-study variation ($I^2 = 38.75\%$), and the *Q* test suggests that the amount of heterogeneity in the effect size estimate is significantly more than would be expected due to chance ($Q = 54.77$, $df = 34$, $p = .01$). We might conclude from the significant *Q* test that heterogeneity in the overall effect size estimate is concerning. At the same time, however, it is not immediately apparent how the *Q* test and the I^2 value relate to the magnitude of heterogeneity in the estimate (τ^2), which has a value of 0.02. Considering these outputs, we are left with an unclear picture of the impact of heterogeneity in the overall effect size estimate, because despite the result of the *Q* test and the I^2 value, we still do not know whether there is minimal or considerable variation in true effects between studies. We do not know this because we have not yet examined the range of true effects across primary studies.

Prediction Intervals

The distribution or range of true effects across primary studies is known as a prediction interval, and is a more direct presentation of the degree to which effects vary between studies than other commonly used heterogeneity indicators (Borenstein et al., 2017; Higgins, 2008). A prediction interval is calculated using the estimated τ^2 value. As mentioned above, the square-root of τ^2 is the standard deviation of true effects across primary studies; it is, therefore, an index of the dispersion of true effects around the overall effect size estimate. When a 95% prediction interval is calculated, it represents the range within which we would expect 95% of all primary studies' true effect sizes to fall (Borenstein et al., 2017; Higgins, Thompson, & Spiegelhalter, 2009). Consequently, a prediction interval offers an intuitive interpretation: A narrow prediction interval indicates minimal variation in true effects between studies, and a wide prediction interval indicates extensive variation in true effects between studies. With this straightforward interpretation, we are better able to answer the key questions related to evaluating the impact of heterogeneity: How variant are effect sizes between studies, and how much concern should be raised by that variation? Put another way: If we conduct a similar study in the future, what is the reasonable range of effects we might expect to achieve?

In the previous section, we illustrated some of the challenges in assessing the impact of between-study variation in true effects using other heterogeneity indicators, and we can now return to the BAI meta-analysis example to show the utility of prediction intervals for this purpose. The overall effect size estimate was −0.21, which is accompanied by a prediction interval of −0.48 to 0.07. With this range, we can make a direct assessment of between-study variation (heterogeneity): In general, the expected effect of a BAI on alcohol-related consequences would be negative (i.e., participants receiving the BAI would have lower post-intervention levels of alcohol-related consequences than participants in comparison conditions), but there is a possibility that some studies may observe higher levels of alcohol-related consequences in the BAI (versus comparison) condition. From this assessment, we can conclude that although there is heterogeneity in effects between studies, the overall estimate is fairly representative of the distribution of true effect sizes among primary studies given that most studies would be expected to have an effect in the same direction as the overall effect size estimate (both negative, or in favor of the BAI).[9]

In the BAI example, the effect size measure is a standardized mean difference; as noted above, the square-root of τ^2 shares the same unit as the effect size measure and, in the case of a standardized mean difference, has a fairly intuitive standalone interpretation as standard deviation-unit differences between studies. We also noted that this straightforward interpretation is not the case when effect size measures such as log odds or risk ratios, prevalences, or Fisher's-z transformed correlations are used, because τ, while still defined as a standard deviation, takes on log or other units that are not as directly interpretable as a standalone value. A benefit of the prediction interval is that the units of the interval can be returned to the unit of the effect size measure (e.g., through exponentiation), so that when, for example, an overall effect size measure is in the form of a log odds ratio, the accompanying prediction interval can express a range of odds ratios.[10] Prediction intervals can also be usefully incorporated into a forest plot, one of the most frequently-used visualizations in meta-analysis. In a forest plot (see Figure 2), the estimated effect size and corresponding confidence interval of each primary study are plotted. Each study's estimated effect size is indicated by a box (aligning with a scale of effect sizes on the x-axis), and the size of the box corresponds to the weight given to each study; studies with greater weight have a larger box. The overall effect size estimate is typically placed at the base of the plot, where it can be visually compared with the individual effect estimates. The utility of the forest plot is that it graphically summarizes the key information of a meta-analysis, namely the study-level and overall effect estimates and the precision with which those effects were estimated; as a result, the plot clearly presents the range of *observed* effects in a meta-analysis. Notably, however, forest plots do not typically display the estimated range of *true* effects, because they do not usually include a prediction interval for the overall effect size

[9]It is important to note the distinction between a confidence interval and a prediction interval as they relate to an overall effect size estimate. The confidence interval concerns the single, overall effect size estimate, and the precision with which it has been estimated. The prediction interval concerns the entire population of true effect sizes, and how they are dispersed about the overall effect size estimate. A wide confidence interval, for instance, tells us that our estimated average effect may be far from the true average effect because we have imprecisely estimated that true average effect. A wide prediction interval, on the other hand, tells us that our estimated average effect may be unrepresentative of the distribution of true effects.

[10]Borenstein et al. (2017) provide methods for calculating error-adjusted prediction intervals for different types of effect sizes, including means, ratios, prevalences, and correlations. R code illustrating the calculation of prediction intervals using the *metafor* package is included in the online supplementary materials to this article.

estimate. When the prediction interval is incorporated, the forest plot presents both sources of variation in the overall effect size estimate – within-study error (the confidence interval for each study's estimated effect) and true between-study heterogeneity – and therefore displays a more informative summary of variation in a meta-analysis. The forest plot in Figure 2, which uses the BAI meta-analysis data, includes a prediction interval as the gray dashed line around the overall effect size estimate (the diamond-shaped indicator). It can be clearly seen that, despite variation in true effects between studies, most study-level effects would be in the same direction as the overall effect size estimate (i.e., in favor of the BAI). Prediction intervals, alone or included in forest plots, are currently underutilized in the field of adolescent development, and we encourage their broader use as an approach for expressing and evaluating heterogeneity in meta-analysis.

Explaining Heterogeneity

In the prior section, we described prediction intervals as an approach to assessing heterogeneity that provides a relatively intuitive and unambiguous means to evaluate how well an overall effect size estimate represents the distribution of true effects among primary studies. A question that naturally follows from a process of assessing heterogeneity is whether sources of heterogeneity can be identified, and in particular, whether those sources of between-study variation are related to aspects of methodological complexity among primary studies. Investigating methodological attributes of primary studies as potential sources of heterogeneity can be informative in several ways. Observing differences in effects based on program characteristics, such as method of delivery, or by aspects of study implementation, such as the duration or modality of training for interventionists, can inform program implementation in both research and applied contexts. Additionally, heterogeneity associated with major design characteristics of primary studies, such as the method of random sequence generation used or whether outcome assessors were blinded, can clarify the role of these factors in influencing overall effect size estimates, and lend support for more rigorous designs.[11] While there are several approaches to examining variation in effects between primary studies in a meta-analysis, in the next section we focus on meta-regression as a flexible and informative tool for this purpose.[12]

Meta-regression

Meta-regression is a versatile framework for investigating the influence of primary study characteristics on effect size heterogeneity in a meta-analysis. Meta-regression is a special case of the general linear model with heteroskedastic sampling variances that are assumed to be known; in this form of linear regression, a linear model can be fitted containing study-level variables (described here as covariates) to assess whether they are related to, or moderate, the magnitude of effect size estimates (Thompson & Higgins, 2002). Meta-

[11]Ideally, potential moderators of study effect sizes should be specified a priori and guided by, for instance, underlying theories of change, input from stakeholders, and/or prior empirical findings. Investigating moderators without a priori hypotheses can be appropriate, but such analyses should be explicitly acknowledged as post hoc and the analyst should consider employing methods to control inflated Type I error associated with multiple tests.

[12]Subgroup analysis is another common approach for examining between-study variation in a meta-analysis and can be applied to groups of studies based on methodological attributes. Subgroup analysis may also incorporate analysis of variance techniques to assess subgroup differences. Further detail on subgroup analysis is provided by Borenstein and Higgins (2013).

regression shares many of the same features as regression in primary research, including the ability to accommodate continuous and categorical covariates, and to include covariates in main effects, polynomials, and/or multiplicative interactions. Meta-regression also bears the same considerations as linear regression in primary research, foremost of which is that the choice and number of covariates included in a model can introduce biases and yield misleading results, such as when important confounding variables have been omitted from the model or if highly collinear variables are included simultaneously in a model. When applied appropriately, however, meta-regression models that include covariates reflecting methodological complexity among primary studies can help to characterize heterogeneity in a number of ways. First, meta-regression models can provide an omnibus test of the significance of included covariates using a null hypothesis which states that none of the covariates are associated with effect size magnitude (Borenstein, Hedges, Higgins, & Rothstein, 2009). In this way, a methodological attribute of primary studies, such as differences in intervention delivery method, can be assessed as a moderator of effect size magnitude. If a covariate reflecting complexity has a significant association with effect size magnitude, meta-regression can be used to further examine how estimated effects vary relative to specific levels of the covariate (e.g., by specific types of intervention delivery methods). Finally, if after identifying a source of between-study variation, there is an interest in examining additional aspects of complexity as potentially contributing to heterogeneity, meta-regression models permit investigating other complexity-related covariates while the influence of the known sources of heterogeneity is held constant (i.e., using those covariates as control variables).

Examining heterogeneity with meta-regression.—Continuous covariates related to methodological complexity that can be incorporated into a meta-regression may be, for example, the duration of an intervention or of interventionist training, or length of time until follow-up assessment. Examples of categorical covariates might include study design (e.g., randomized controlled vs. quasi-controlled), comparison condition type (e.g., no treatment vs. treatment as usual), or risk of bias ratings (e.g., high vs. low risk of bias associated with outcome assessor blinding). In the BAI meta-analysis example, we may wish to examine the role of intervention delivery method as a source of effect size heterogeneity, as suggested above. Table 1 provides the output for a mixed-effects meta-regression model, where the outcome of interest is post-intervention differences in alcohol-related problems measured as the standardized mean difference (Cohen's d) between intervention and comparison conditions, and the moderator (covariate) tested is intervention delivery method. In each of the primary studies, the BAI was delivered face-to-face, using a pen-and-paper modality, via a computer interface, or using a mixture of delivery methods, and these delivery methods are included in the meta-regression model using three dummy-coded indicator variables.

In the model output, we can first inspect the omnibus test ("Test of Moderators"), which is statistically significant at the $\alpha = .05$ level, suggesting that delivery method does moderate the overall effect size estimate. We can also note two other elements of the output that are informative about delivery method as a source of heterogeneity. The first is the estimated value of τ^2, which now reflects *residual heterogeneity* after accounting for the covariates in the model. As noted above, the estimate of τ^2 was 0.02 before accounting for any study-level

factors that might explain heterogeneity; in the current model, when we control for the effect of delivery method, the estimate of τ^2 is reduced to 0.005. A second useful value in the model output is the amount of heterogeneity explained, or R^2 (Thompson & Higgins, 2002). The R^2 value indicates that delivery method explains about 74% of the true between-study variation. Taken together, these three pieces of information – the significant omnibus test, the substantial reduction in the estimate of τ^2, and the large proportion of explained between-study variation – strongly suggest that delivery method should be viewed as a source of considerable heterogeneity in the overall effect size estimate, one that is related to methodological complexity.

Following such an assessment, we may have an interest in examining differences in the average effect size estimate across different types of delivery methods, and we can do so using the regression coefficients provided in Table 1. In this output, the intercept represents the overall effect size estimate among primary studies that used a face-to-face method for BAI delivery (the reference category for the dummy variable indicators). In these studies, RAPI scores of alcohol-related problems were on average lower $\bar{d} = -0.29$, 95% CI [−0.36, −0.21]) in the intervention versus comparison conditions. The coefficients for other delivery methods are interpreted relative to the reference category of face-to-face delivery method. If we examine the computerized delivery coefficient estimate, we can see that among studies that used this delivery method, the average standardized mean difference between intervention and comparison conditions is substantially smaller compared with face-to-face delivery ($\bar{d} = -0.29 + 0.25 = -0.04$, 95% CI [−0.16, 0.08]). These findings suggest that delivery method is related to whether a BAI is associated with lower levels of alcohol-related problems, and that computerized delivery of BAIs may be associated with smaller beneficial effects on this type of outcome.

Additional considerations.—Employing regression analysis in meta-analytic contexts is accompanied by several considerations beyond those involved in the use of regression in primary research noted above. For one, meta-regression does not preserve primary study-level randomization, and therefore does not permit causal inference about the effects of moderators (Thompson & Higgins, 2002). Consequently, in the example interpretation provided above, we are limited to observational associations and cannot conclude unequivocally that face-to-face delivery of a BAI is more effective in reducing alcohol-related problems compared with computerized delivery. Additionally, meta-regression models are also susceptible to imbalance because the number of studies – and subgroups of studies within values of covariates – is often small. For example, in our example model output, we would exhibit caution in interpreting the coefficient representing the pen-and-paper delivery method, despite its statistical significance: As suggested by its relatively large standard error, few studies used this method in comparison to other modalities (only 2 studies used pen-and-paper delivery, in contrast to 9 using computerized delivery, and 23 employing a face-to-face approach).[13] Furthermore, use of covariates in meta-regression that represent aggregations of study-level characteristics (e.g., demographic composition, such as

[13]Relatedly, heterogeneity may be present within levels of a moderator (e.g., the true effect of the face-to-face delivery method may vary across studies). Applications of multilevel and structural equation modeling for meta-analysis can facilitate examination of such subgroup moderation effects; see Schoemann (2016).

proportion female) may lead to the ecological fallacy, or the incorrect overgeneralization of findings from the aggregate level to the individual participant level (Baker et al., 2009; Berlin, Santanna, Schmid, Szczech, & Feldman, 2002; Thompson & Higgins, 2002).[14] In view of these limitations, meta-regression is often seen as hypothesis generating rather than as a means of hypothesis testing (Thompson & Higgins, 2002).[15] When meta-regression is used to examine heterogeneity in true effects as a result of methodological complexity, findings and resulting research questions can be useful for further study of how specific methodological factors influence the effectiveness of a program.

Conclusion.

By discussing various heterogeneity statistics, visualizations, and explanatory methods, this article provides an overview of frequently employed and useful tools for assessing and explaining heterogeneity in a meta-analysis. There are other statistical and graphical approaches helpful for examining complexity and heterogeneity in meta-analyses, but we have focused on the most commonly used tools. With prediction intervals, by contrast, we have presented a currently underutilized approach. Readers interested in exploring more comprehensive treatments of options to explore heterogeneity are encouraged to examine the primary sources noted in the citations and footnotes included throughout the article. Further, the supplementary materials of this article include the full R code and data used to produce the example BAI meta-analysis results and presentations of heterogeneity we have discussed.

Our aim was also to orient the applied adolescent development researcher to examining methodological complexity as a potential source of heterogeneity in meta-analytic effect estimates. We argued that heterogeneity should be viewed as a key parameter of interest in a meta-analysis, one that can provide valuable information on how effect size estimates may vary by participant, context, and other study characteristics. Indeed, because meta-analysis can be used to synthesize evidence from a large body of primary research evidence, methods such as meta-regression can be useful for examining how a broad array of intervention ingredients or approaches may relate to variation in true effects across studies. Although methods such as meta-regression do not provide causal evidence, they can provide valuable correlational evidence that can guide future research and ultimately enhance the effectiveness of programs aimed at promoting healthy adolescent development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

[14]Overgeneralization in this way would be described as an aggregation bias. To avoid this bias, individual participant data from primary studies, when available, should be used in regression analyses. See Footnote 1 for suggested readings on conducting meta-analysis with individual participant data.

[15]For detailed discussions of considerations – and cautions – in conducting meta-regression, see Thompson and Higgins (2002) and Baker et al. (2009).

Acknowledgements:

The authors wish to thank Dr. Michael Borenstein for his useful comments and suggestions on an earlier version of this article.

Funding: This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The example data reported in this work were originally collected via support from award number R01AA020286 from the National Institute on Alcohol Abuse and Alcoholism. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health.

References

Anzures-Cabrera J, & Higgins JPT (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. Research Synthesis Methods, 1, 66–80. [PubMed: 26056093]

Baker WL, White CM, Cappelleri JC, Kluger J, & Coleman CI (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. International Journal of Clinical Practice, 63(10), 1426–1434. [PubMed: 19769699]

Baujat B, Mahe C, Pignon J-P, & Hill C (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. Statistics in Medicine, 21(18), 2641–2652. 10.1002/sim.1221 [PubMed: 12228882]

Berlin JA, Santanna J, Schmid CH, Szczech LA, & Feldman HI (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. Statistics in Medicine, 21, 371–387. 10.1002/sim.1023 [PubMed: 11813224]

Borenstein M (2018). Common mistakes in meta-analysis and how to avoid them. Englewood, NJ: Biostat, Inc.

Borenstein M, Hedges LV, Higgins JPT, & Rothstein HR (2009). Introduction to meta-analysis. Chichester, UK: Wiley.

Borenstein M, Hedges LV, Higgins JPT, & Rothstein HR (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1, 97–111. 10.1002/jrsm.12 [PubMed: 26061376]

Borenstein M, & Higgins JPT (2013). Meta-analysis and subgroups. Prevention Science, 14, 134–143. 10.1007/s11121-013-0377-7 [PubMed: 23479191]

Borenstein M, Higgins JPT, Hedges LV, & Rothstein HR (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Research Synthesis Methods. Retrieved from http://doi.wiley.com/10.1002/jrsm.1230

Burns MK (Ed.). (2012). Meta-analysis of single-case design research [Special issue] Journal of Behavioral Education, 21(3). Retrieved from https://link.springer.com/journal/10864/21/3/page/1

Card NA, & Little TD (2006). Proactive and reactive aggression in childhood and adolescence: A meta-analysis of differential relations with psychosocial adjustment. International Journal of Behavioral Development, 30(5), 466–480. 10.1177/0165025406071904

Clarke AT (2006). Coping with interpersonal stress and psychosocial health among children and adolescents: A meta-analysis. Journal of Youth and Adolescence, 35(1), 10–23. 10.1007/s10964-005-9001-x

Cooper H, Hedges LV, & Valentine JC (Eds.). (2019). The handbook of research synthesis and meta-analysis (Third). New York, NY: Russell Sage Foundation.

Cooper H, & Patall EA (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychological Methods, 14(2), 165–176. 10.1037/a0015565 [PubMed: 19485627]

Deeks J, Bossuyt P, & Gatsonis C (Eds.). (2013). Cochrane handbook for systematic reviews of diagnostic test accuracy (1.0.0). Retrieved from https://methods.cochrane.org/sdt/handbook-dta-reviews

Durlak JA, Weissberg RP, & Pachan M (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. American Journal of Community Psychology, 45, 294–309. 10.1007/s10464-010-9300-6 [PubMed: 20300825]

Hardy RJ, & Thompson SG (1998). Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine, 17, 841–856. [PubMed: 9595615]

Hedges LV (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92(2), 490–499. 10.1037/0033-2909.92.2.490

Hedges LV, & Vevea JL (1998). Fixed- and random-effects models in meta-analysis. Psychological Methods, 3(4), 486–504.

Higgins JPT (2008). Heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology, 37, 1158–1160. [PubMed: 18832388]

Higgins JPT, & Green S (Eds.). (2011). Cochrane handbook for systematic reviews of interventions. Retrieved from http://handbook.cochrane.org

Higgins JPT, & Thompson SG (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21, 1539–1558. [PubMed: 12111919]

Higgins JPT, Thompson SG, & Spiegelhalter DJ (2009). A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(1), 137–159. 10.1111/j.1467-985X.2008.00552.x

Hoaglin DC (2015). We know less than we should about methods of meta-analysis. Research Synthesis Methods, 6, 287–289. 10.1002/jrsm.1146 [PubMed: 26096892]

Horowitz JL, & Garber J (2006). The prevention of depressive symptoms in children and adolescents: A meta-analytic review. Journal of Consulting and Clinical Psychology, 74(3), 401–415. 10.1037/0022-006X.74.3.401 [PubMed: 16822098]

Kulinskaya E, Dollinger MB, & BjØrkestØl K (2011). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference: Cochran's Q for risk difference. Research Synthesis Methods, 2, 254–270. 10.1002/jrsm.54 [PubMed: 26061889]

Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, … Simmonds M (2018). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods, 1–16.

Langan D, Higgins JPT, & Simmonds M (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. Research Synthesis Methods, 8, 181–198. 10.1002/jrsm.1198 [PubMed: 27060925]

Lipsey MW, & Wilson DB (2001). Practical meta-analysis. Thousand Oaks, CA: Sage Publication, Inc.

Lorenc T, Felix L, Petticrew M, Melendez-Torres GJ, Thomas J, Thomas S, …Richardson M (2016). Meta-analysis, complexity, and heterogeneity: A qualitative interview study of researchers' methodological values and practices. Systematic Reviews, 5(192). 10.1186/s13643-016-0366-6

Mittlbock M, & Heinzl H (2006). A simulation study comparing properties of heterogeneity measures in meta-analyses. Statistics in Medicine, 25, 4321–4333. [PubMed: 16991104]

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, & Stroup DF (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. The Lancet, 354, 1896–1900.

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1). Retrieved from https://www.R-project.org/

Riley RD, Higgins JPT, & Deeks JJ (2011). Interpretation of random effects meta-analyses. BMJ, 342 10.1136/bmj.d549

Riley RD, Lambert PC, & Abo-Zaid G (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. BMJ, 340, c221 10.1136/bmj.c221 [PubMed: 20139215]

RStudio Team. (2018). RStudio: Integrated development environment for R (Version 1.2.1335). Retrieved from http://www.rstudio.com/

Schoemann AM (2016). Using multiple group modeling to test moderators in meta-analysis. Research Synthesis Methods, 7, 387–401. http://doi.wiley.com/10.1002/jrsm.1200 [PubMed: 27936303]

Shadish WR, Hedges LV, & Pustejovsky JE (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. Journal of School Psychology, 52(2), 123–147. 10.1016/j.jsp.2013.11.005 [PubMed: 24606972]

Stewart LA, Tierney JF, & Clarke M (2011). Reviews of individual patient data In Higgins JPT & Green S (Eds.), Cochrane handbook for systematic reviews of interventions (5.1.0). Retrieved from http://handbook.cochrane.org

StrØm HK, Adolfsen F, Fossum S, Kaiser S, & Martinussen M (2014). Effectiveness of school-based preventive interventions on adolescent alcohol use: A meta-analysis of randomized controlled trials. Substance Abuse Treatment, Prevention, and Policy, 9(48). 10.1186/1747-597X-9-48

Tanner-Smith EE, & Grant S (2018). Meta-analysis of complex interventions. Annual Review of Public Health, 39, 135–151.

Tanner-Smith EE, & Lipsey MW (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. Journal of Substance Abuse Treatment, 51, 1–18. [PubMed: 25300577]

Tanner-Smith EE, & Risser MD (2016). A meta-analysis of brief alcohol interventions for adolescents and young adults: Variability in effects across alcohol measures. The American Journal of Drug and Alcohol Abuse, 42(2), 140–151. [PubMed: 26905387]

Tanner-Smith EE, Wilson SJ, & Lipsey MW (2013). Risk factors and crime In Cullen FT & Wilcox P (Eds.). The Oxford handbook of criminological theory (pp. 89–111). New York: Oxford University Press.

Thompson SG (1994). Why sources of heterogeneity in meta-analysis should be investigated. BMJ, 309(6965), 1351–1355. 10.1136/bmj.309.6965.1351 [PubMed: 7866085]

Thompson SG, & Higgins JPT (2002). How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine, 21, 1559–1573. 10.1002/sim.1187 [PubMed: 12111920]

Tripodi SJ, Bender K, Litschge C, & Vaughn MG (2010). Interventions for reducing adolescent alcohol abuse: A meta-analytic review. Archives of Pediatrics & Adolescent Medicine, 164(1), 85–91. 10.1001/archpediatrics.2009.235 [PubMed: 20048247]

Valentine JC, Konstantopoulos S, & Goldrick-Rab S (2017). What happens to students placed into developmental education? A meta-analysis of regression discontinuity studies. Review of Educational Research, 87(4), 806–833. 10.3102/0034654317709237

Valentine JC, Tanner-Smith EE, Pustejovsky JE, & Lau TS (2016). Between-case standardized mean difference effect sizes for single-case designs: A primer and tutorial using the scdhlm web application. Retrieved from https://campbellcollaboration.org/media/k2/attachments/1_Effect_sizes_for_single_case_designs.pdf

Viechtbauer W (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics, 30(3), 261–293.

Viechtbauer W (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. Zeitschrif Für Psychologie / Journal of Psychology, 215(2), 104–121.

Viechtbauer W (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3). 10.18637/jss.v036.i03

Viechtbauer W (2017). Package “metafor” (Version 2.0-0). Retrieved from https://cran.r-project.org/web/packages/metafor/index.html

Weisz JR, McCarty CA, & Valeri SM (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. Psychological Bulletin, 132(1), 132–149. 10.1037/0033-2909.132.1.132 [PubMed: 16435960]

White HR, & Labouvie EW (1989). Toward the assessment of adolescent problem drinking. Journal of Studies on Alcohol, 50, 30–37. 10.15288/jsa.1989.50.30 [PubMed: 2927120]

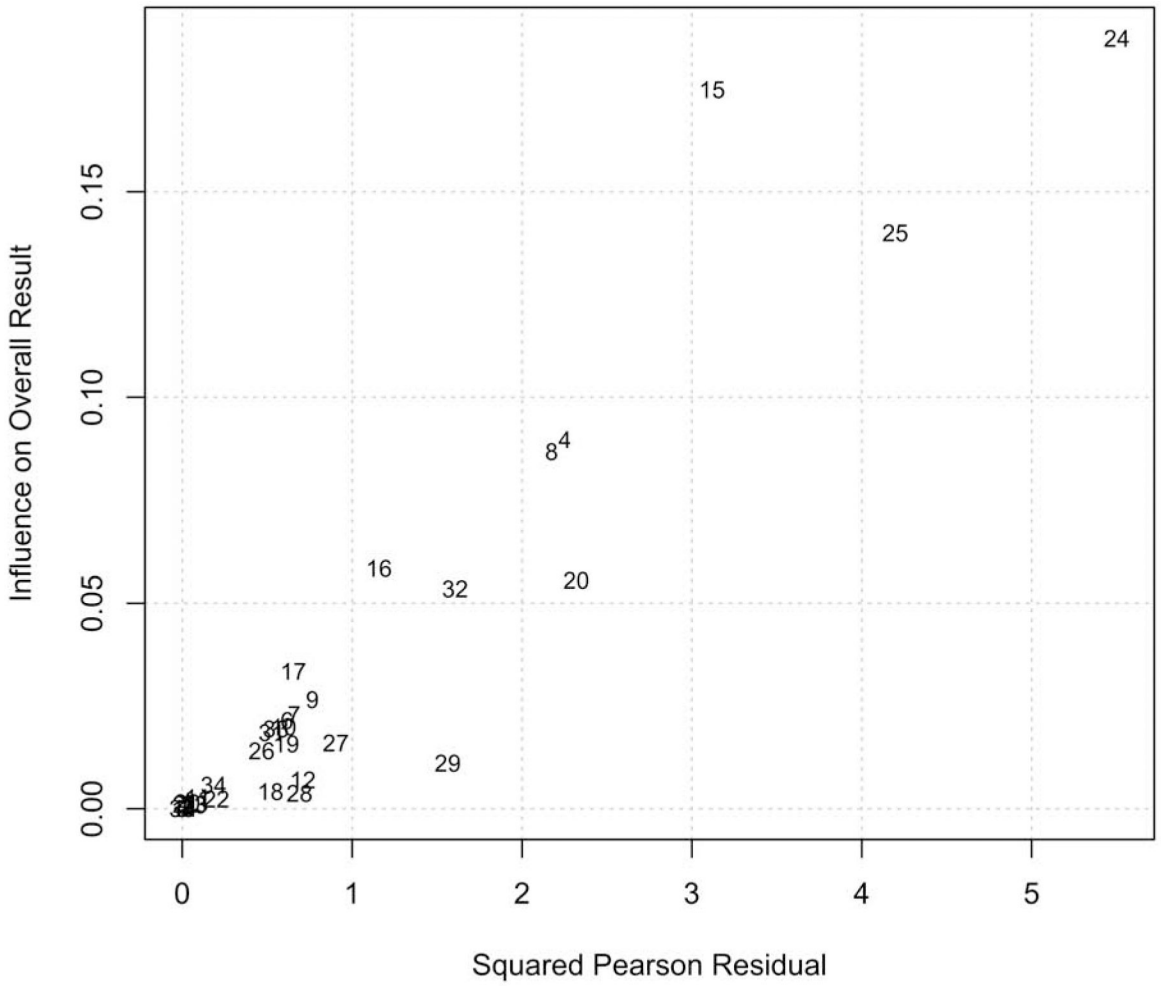

Figure 1.

Baujat plot using the BAI random-effects model. The x-axis is the squared Pearson residual of each study; a larger residual suggests a study's estimated effect is outlying from the overall effect estimated by the model. The y-axis corresponds to the standardized squared difference between the model-estimated effect for each study, with and without each study included in the model estimation. Primary studies with the greatest variation from the overall effect size estimate and the most substantial contribution to the estimate are located in the upper-right corner of the plot.

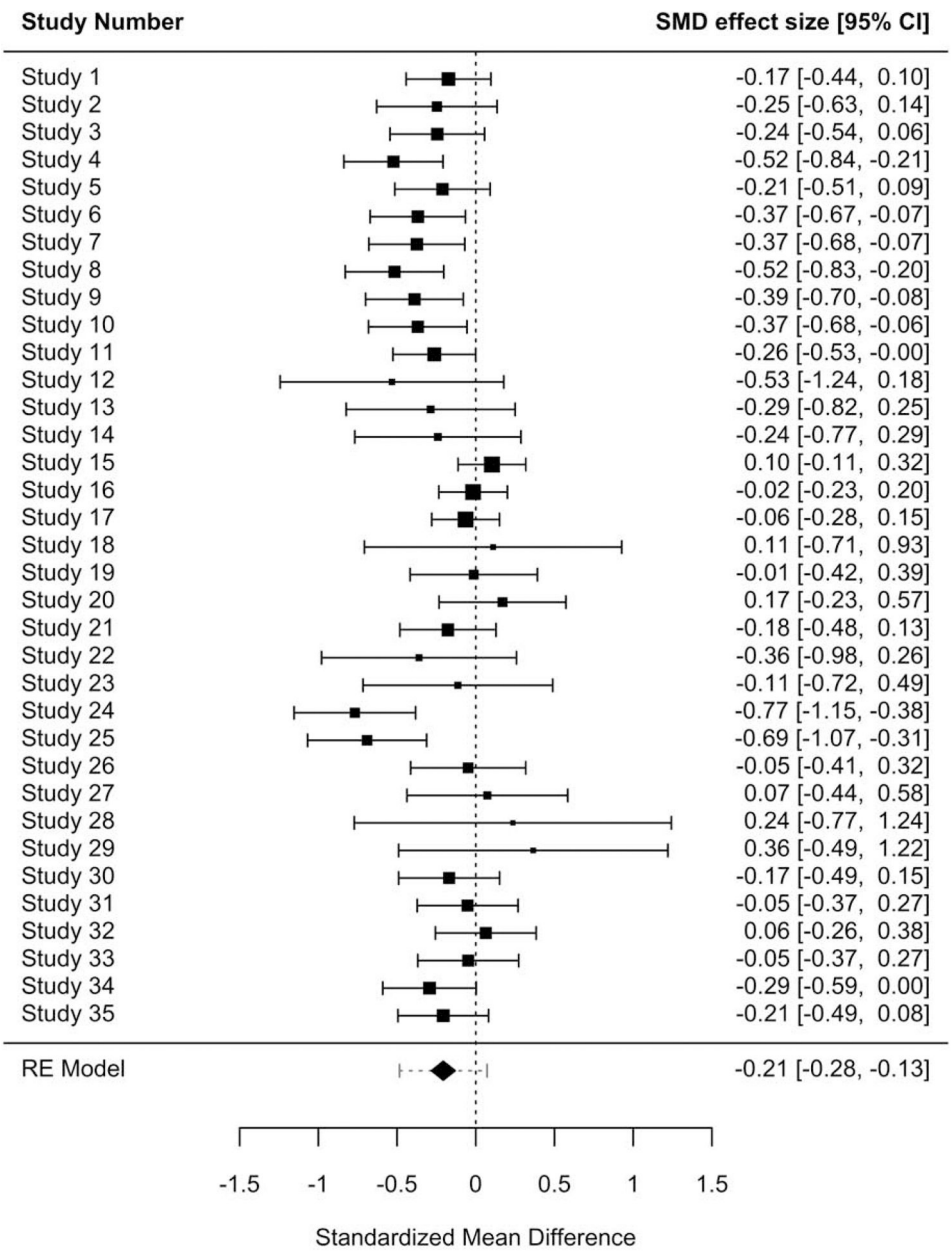

Figure 2.

Forest plot of BAI meta-analysis primary study effect size estimates with prediction interval. The overall effect size estimate is indicated by the diamond-shaped indictor at the bottom of the plot; its width corresponds to the 95% confidence interval for the estimate, which is also provided in brackets on the same line on the plot. The 95% prediction interval is denoted by gray bounding lines around the overall effect size estimate indicator.

Note. SMD = standardized mean difference. CI = confidence interval. RE = random-effects.

Table 1

Mixed-effects Meta-regression Model of Post-Intervention Differences in Alcohol-Related Problems Moderated by Intervention Delivery Method (k = 35)

BAI Delivery Method	*b*	SE	95% CI	
Intercept	−0.29***	0.04	[−0.36, −0.21]	
Pen-and-paper	0.36*	0.16	[0.05, 0.68]	
Computerized	0.25**	0.07	[0.11, 0.39]	
Mixed modes	−0.25	0.36	[−0.97, 0.48]	
τ^2				0.005
R^2				73.75
Test of Moderators				$F(3, 31) = 5.65$, $p = .003$

Note. Effect size is standardized mean difference between intervention and comparison conditions. In this output, the intercept is the coefficient for the face-to-face delivery method. *k* = number of primary studies. BAI = brief alcohol intervention. *b* = unstandardized model coefficient. SE = standard error. CI = confidence interval.

* $p < .05$.

** $p < .01$.

*** $p < .001$.