Methods for Synthesizing Findings on Moderation Effects Across Multiple Randomized Trials

(Article begins on next page)

18 July 2024

This is the author's final version of the contribution published as:

The publisher's version is available at:
http://link.springer.com/content/pdf/10.1007/s11121-011-0207-8

When citing, please refer to the published version.

Link to this full text:
http://hdl.handle.net/2318/83677

Methods for Synthesizing Findings on Moderation Effects Across Multiple Randomized Trials

C Hendricks Brown, University of Miami
Zili Sloboda, JBS International
Fabrizio Faggiano, Avogadro University
Brent Teasdale, Georgia State University
Ferdinand Keller, University of Ulm
Gregor Burkhart, European Monitoring Centre for Drugs and Drug Addiction
Federica Vigna-Taglianti, Piedmont Centre for Drug Addiction Epidemiology
George Howe, George Washington University
Katherine Masyn, Harvard University
Wei Wang, University of South Florida
Bengt Muthén, UCLA
Peggy Stephens, Akron University
Scott Grey, Kent State University
Tatiana Perrino, University of Miami

Prevention Science and Methodology Group

Abstract

This paper presents new methods for synthesizing results from subgroup and moderation analyses across different randomized trials. We demonstrate that such a synthesis generally results in additional power to detect significant moderation findings above what one would find in a single trial. Three general methods for conducting synthesis analyses are discussed, with two methods, integrative data analysis, and parallel analyses, sharing a large advantage over traditional methods available in meta-analysis. We present a broad class of analytic models to examine moderation effects across trials that can be used to assess their overall effect and explain sources of heterogeneity, and present ways to disentangle differences across trials due to individual differences, contextual level differences, intervention, and trial design.

Introduction

Through the use of meta-analysis(Durlak & Wells, 1997; Faggiano et al., 2005; Faggiano et al., 2008; Tobler, 1986) and scientific reviews(Elliott, Mihalic, Elliott, & Mihalic, 2004; O'Connell, Boat, & Warner, 2009) that are applied to findings from randomized trials testing specific interventions, we now have identified a large number of programs or interventions that have been shown to be efficacious or effective in the prevention of mental disorders, drug abuse, and delinquency. A set of generally accepted procedures has emerged to guide the searching for trials, coding of trial results, steps in conducting meta-analysis, and summarization of evidence (Higgens & Green, 2008). Generally, the three major dimensions used to make these decisions about evidence are based on 1) the statistical magnitude of overall impact on a targeted behavioral outcome, 2) determining whether the trial was designed and conducted with sufficient quality to support causal conclusions, and 3) evaluating the replicability of findings across multiple trials of the same or similar intervention. Such information has served as the basis for selecting evidence-based programs for wide scale dissemination and implementation (Brown, Hawkins, Arthur, Briney, & Abbott, 2007) , although different review groups vary in the criteria they require interventions to meet across these three dimensions (Flay et al., 2005).

One limitation of this current system for determining which programs meet standards of evidence is that it does not account for how programs may benefit, or potentially harm, subgroups within a defined population. A related limitation is that these evaluations do not describe whether or to what extent intervention effects vary across contexts. There are numerous examples of interventions that affect subjects differently based on their own baseline risk (Brown & Liao, 1999; Brown et al., 2008; Tein, Sandler, MacKinnon, & Wolchik, 2004; Wolchik, Sandler, Weiss, & Winslow, 2007), that produce iatrogenic effects with adolescents who learn deviant behavior (Dishion, Burraston, & Poulin, 2001; Dishion, McCord, & Poulin, 1999; Dishion, Spracklen, Andrews, & Patterson, 1996), or that show differential effects across community settings (Van Horn et al., under review, 2008). Such knowledge would be valuable in identifying interventions that address differing risk and protective factors, mediational pathways, cultural factors, or community preferences and resources. Traditional summarizations of evidence need extending to allow delivery of interventions that match community and individual needs, preferences, and likelihood of receiving benefit.

In this paper, we provide a methodologic perspective on how to use multiple randomized trials to understand how an intervention's effect varies or remains constant across individuals and contexts. We show that single trials are generally underpowered to examine variation in impact, and combining data across multiple randomized trials can increase statistical power for modeling variation in impact. Combining such moderation data, however, requires new analytic models for synthesizing findings, different ways of decomposing sources of variation across trials, and alternative ways of combining data based on the degree that data can be shared.

We use the broad term of "variation in impact" to refer to individual or contextual factors at baseline that affect the relationship between intervention and outcome. These sources can be measured quantities or attributes, such as age or gender, or unmeasured characteristics, whose presence can only be inferred indirectly by identifying a significant source of variation in impact through mixture models or multilevel modeling. The term "moderation" will refer only to variation in impact through measured baseline variables. These moderators can be at the individual level, such as age or ethnicity, or across measured cultural or contextual factors, such as neighborhood rates of underage drinking. Moderation is generally tested with interaction terms involving a covariate and intervention status. The term "heterogeneity" will refer to sources of variation that are present but not clearly identified. For example, when an intervention's impact varies significantly across a set of trials or the interaction of gender and intervention status on outcome varies

significantly across trials, we refer to either of these as displaying heterogeneity of effects. In this paper, we use the specialized term "subgroup analyses" to refer to a restricted set of analyses where intervention effect is examined only within a subset of the sample, i.e., males, and no attempt is made to assess the comparative effect of the intervention across different subsets.

This paper addresses several important methodologic challenges in identifying and quantifying moderation effects of interventions. The foremost challenge is that moderation effects are difficult to assess because these baseline by intervention interaction analyses are very often underpowered in a single randomized trial. We first determine conditions under which there are gains in power when using data from multiple trials compared to a single trial. Our development is based on short statistical arguments for the general reader; footnotes buttress these short presentations with more details on these statistical arguments. We then present a new, general multilevel approach for decomposing moderator effects both within and between trials. These analytic models for synthesis of moderator effects are only useful when two problems can be solved. First, we need to be able to distinguish functionally different reasons for variation within and across trials. Do the outcomes from two trials differ from underlying differences in growth patterns, or merely from different measurement times for outcomes in the two studies? Our approach involves growth modeling, and in particular multilevel growth mixture modeling to calibrate change across trials with different follow-up times. The second challenge in using these models is that they require access to data of sufficient depth within each trial and sufficient breadth of data across trials to carry out these types of moderator analyses. We present three alternative approaches to combining data across trials based on the level of data sharing available, and we compare their strengths and weaknesses. In our conclusion we give guidance on when to use these methods and their limitations.

What Do We Want to Achieve from Moderator Analyses?

Often embedded in many etiologic theories are specific hypotheses about moderation. Our theory of change that underlies how we believe an intervention should work often leads to a priori moderation hypotheses. For prevention science, the fundamental paradigm involves identifying antecedent risk and protective factors leading towards a target outcome, then applying an intervention to interrupt the risk process or strengthen protective factors (Coie, Watt, West, & Hawkins, 1993; Howe, Reiss, & Yuh, 2002; Kellam & Langevin, 2003). This general framework suggests examining the degree to which risk or protective factors moderate an intervention's effect. For universal interventions that target early risk behaviors within a developmental epidemiologic perspective, which normative systems such as classrooms and schools to reinforce prosocial behavior, we often predict that the most benefit will occur among those with an expressed risk factor at baseline, (Brown, Wang, Kellam, et al., 2008; Dolan et al., 1993; Ialongo, Poduska, Werthamer, & Kellam, 2001; Ialongo et al., 1999; Kellam et al., 2008; Kellam, Koretz, & Moscicki, 1999).

We are often interested in examining the preventive effects on low and high risk youth separately. Thus a middle school-based drug prevention program may have different effects on those who already use substances at baseline versus those who do not. An intervention designed primarily to address only one of these subgroups, say to prevent initiation, may have negative effects on the other subgroup of users. In fact, one of the criticisms of the original DARE program, was that the delivery of the program by police officers might alienate those youth who were already engaged in deviant behavior (Ennett, Tobler, Ringwalt, & Flewelling, 1994). In a recent trial that used DARE officers with an updated curriculum, such early deviant youth were more engaged, but this program may have inadvertently heightened later drug experimentation among those who did not use substances at baseline (Sloboda et al., 2009).

Power to Study Moderation of Intervention Effects

Most trials are powered to detect main effects, so we briefly discuss how the power for moderation analysis relates to that for main effects. Comparison of power hinges on the comparison of standard errors for main effect and moderation estimators[1]. Consider testing for a main effect of intervention with traditional error rates ($\alpha = 0.05$, $\beta = 0.2$) and two-sided testing. For a continuous outcome with equal numbers of individuals assigned to intervention or control, one needs 126 total subjects when the standardized mean difference or effect size is large (ES = 0.5) and 350 subjects when the effect size is more modest (ES =0.3; Cohen, 1988). The test statistic compares the difference in sample means for treatment (t) and control (c), $\bar{X}t - \bar{X}c$ to the main effect (ME) standard error,

$$se_{ME} \ (Individual \ Randomized \ Trial) = 2\widehat{\sigma} / \sqrt{N} \quad (1)$$

where N is the total sample size and $\widehat{\sigma}$ is the standard deviation estimate.

For moderator or interaction effects involving a binary baseline measure, say gender, we would compare the mean differences in intervention effect for males, $\bar{X}tm - \bar{X}cm$, to that for females, $\bar{X}tf - \bar{X}cf$, where the second subscript refers to gender. The standard error of this interaction (Int) ES, depends on the proportion of males, p;

$$se_{int} \ (Individual \ Randomized \ Trial) = 2\widehat{\sigma} / \sqrt{p \, (1 - p) \, N}. \quad (2)$$

A comparison of (1) and (2) shows that the standard error for the interaction term is larger by a factor of . For all possible values of the proportion of males, this factor is always larger than 2, and since power depends inversely on the square of the standard error, one would need at least 4 times the sample size to achieve the same statistical power for testing an interaction that has the same ES as that for a main effect. That means, for an interactive ES of 0.5 the sample size would need to be at least 504 rather than 126, and for an interactive ES of 0.3, the sample size would need to exceed 1400 rather than 350. If the proportion who are in the subgroup is far from 1/2, this would require much more than 4 times the sample size as that for the main effect analysis.

[1]Our argument below provides a partial justification due to space; the complete proof involves formulas for power based on noncentrality parameters, which in turn depend on sample size.

Statistical Power for Testing Moderator Effects in Group Based Trials

In group-based randomized trials, moderator analyses lose less power compared to main effect analyses. Consider conducting a group randomized trial, say when intervention is assigned at the school, classroom, or community level, and we are examining an individual level baseline variable, such as gender, for its moderating effect. In this case, the standard error depends in a more complex way on the number of groups or units that are randomized (M), the number of subjects within each unit (N), all of whom receive the same intervention condition, and the two sources of variance, between (b) and within (w) groups.

$$se_{ME} \ (Group \ Randomized \ Trial) = \sqrt{4\frac{\widehat{\sigma_B^2}}{M} + 2\frac{\widehat{\sigma_W^2}}{N}}$$

$$se_{Int} \ (Group \ Randomized \ Trial) = \sqrt{4\frac{\widehat{\sigma_B^2}}{M} + 2\frac{\widehat{\sigma_W^2}}{p(1-p)N}}.$$

Note that only the second term in these expressions is changed when we turn to tests of interaction. In school-based randomized trials where the number of units M is relatively small, the number of subjects per unit is moderate or large, and the intraclass correlation (ICC) is fairly large, changes in

this second term have less effect on the standard error. As an example, with M = 12 schools, N = 200 subjects per school, and an ICC of 0.05, the standard error for the interaction is 40% larger than that for the main effect, compared to 100% larger for an individual based trial. As we increase the number of subjects per school, the power for testing interactions in group randomized designs approaches that for testing a main effect. A similar situation occurs when there is randomization within blocks, such as a school, where subjects within the same school are assigned to either intervention or control conditions (Brown & Liao, 1999). This situation is analogous to combining multiple randomized trials, where each trial forms a block and there are both intervention and control subjects in each block; consequently this case is covered below under our discussion of integrative data analysis below.

Because sample sizes for trials are almost universally based on detecting significant main effects, few of these trials have any real hope of finding significant moderator effects when they are small to moderate[2]. This basic result has pushed us to consider more powerful moderating analyses involving multiple randomized trials.

[2]The development in this part of the text is limited to interactions involving a binary covariate. The power for detecting a linear interaction with a continuous baseline measure can be compared to that of the main effect once a common calibration of "effect size" is established. Our choice is to scale the treatment variable to have the same variance as that of the continuous variable. The regression coefficient of the interaction term measures the difference in response under intervention and control for two covariate values separated by 1 standard deviation, i.e. $ES_{Inter} = E(Y \mid T = 1, X = 1) - E(Y \mid T = 0, X = 1) - \{E(Y \mid T = 1, X = 0) - E(Y \mid T = 0, X = 0)\}$. To achieve the same power for detecting an effect size , ES_{ME} for the main effect in a trial with equal allocations to intervention and control, we require $ES_{Inter} = 2\ ES_{ME}$. This is the identical result for the case of a dichotomous moderator variable presented in the text.

When Can the Use of Multiple Randomized Trials Increase Statistical Power for Moderator Analyses?

We next examine conditions under which combining data from multiple trials increases power to detect moderation. Our approach to this problem is to model the interactions in each of the j trials in a hierarchical fashion. At the first level, let the individual level response Y_{ij} of subject i within trial j, depend on the same covariate X_{ij} and treatment condition T_{ij}, and their interaction term, $X_{ij}T_{ij}$ representing a moderating effect, with separate coefficients for each trial,

$$Y_{ij} = a_j + a_j^0 X_{ij} + a_j^1 T_{ij} + b_j X_{ij} T_{ij} + \varepsilon_{ij}. \quad (3)$$

Here the last term expresses individual level error with mean 0 and within trial variance .
At the second level we assume that the moderator effect for the jth trial, b_j has a normal distribution with mean b and variance σ_b^2. For a single trial based on N subjects equally allocated to intervention or control, the standardized (Var(X) = 1) estimate $\hat{b}_j$ of the interaction has variance $\dfrac{4\sigma_w^2}{N Var(x)}$
The estimator for the common interaction effect b, obtained from a two-level analysis involving M trials, has variance

$$\frac{\sigma_b^2}{M} + \frac{4\sigma_w^2}{NM Var(x)}$$

Thus the precision of the two-level estimator that synthesizes the findings from multiple trials will be higher than that for a single trial whenever

$$\sigma_b^2 / \sigma_w^2 < \frac{4(M-1)}{N}. \quad (4)$$

In words, the left hand side of this inequality compares the between trial variance to within trial variance while the right hand side depends only on the number of trials and subjects per trial. A synthesis will have increased precision over that of a single trial when the between variance is small, or the number of trials relative to number of subjects is large. A quick way to compare this is based

on the size of the intraclass correlation (ICC), $\sigma_b^2 / \left(\sigma_b^2 + \sigma_w^2 \right)$, applied to Equation (4).

As long as the ICC is less than $\frac{4(M-1)}{N}$. we are guaranteed to increase precision for moderation by combining results across trials[3]. It is rare for ICC's to exceed 5%, so combining results from just two trials is nearly certain to increase power if there are 80 or fewer subjects in a trial. For 4 trials a synthesis will increase power for moderation as long as there are less than 240 subjects per trial. Thus most often the combination of even a small number of trials is likely to provide gains in precision over that in a single trial.

[3]In this argument we have ignored the differences in smaller degrees of freedom needed to test for this interaction effect across trials; nevertheless, the relationship ICC < 4(M−1)/N is still a very conservative bound.

Modeling of Variation in Impact Within and Across Randomized Trials

Any synthesis of intervention findings across trials needs to identify the sources of variation that can be explained by covariates as well as those remaining unexplained. We recommend a synthesis approach that not only obtains a combined overall estimate of a moderator effect but also examines alternative sources of heterogeneity in these moderating effects both within and across trials. In this section we present general modeling approaches using the notation of latent classes and covariates in two-level modeling. To make these ideas concrete, we consider combining data from multiple trials of a school-based intervention for preventing drug abuse. Each trial randomly assigns schools to receive the intervention or serve as a control. One can consider the moderator as individual level smoking at baseline and the outcome as the frequency of marijuana use at follow-up. Table 1 presents a set of general models that can distinguish shared versus unique moderating effects (of baseline smoking) that may be explained by measured or unmeasured covariates at the trial level. For simplicity, we ignore levels of clustering, such as the classroom or school, as well as multiple outcomes in this table; these factors can be added to our models, but they do not introduce any new concepts. All these models in Table 1 assume that the moderating effect is the same for everyone within each trial. Table 2 extends these models by allowing the moderating effects (of baseline smoking) within trials to vary by measured or unmeasured variables as well. We use bold font for emphasis.

The first column in Table 1 is built on the same two-level model as in Equation 1, with trial level random effects for intercept a_j, covariate (baseline smoking status) a_j^0, treatment main effects a_j^1 and treatment by (baseline smoking) covariate moderating effects b_j. All four of these random effects may not be needed in every synthesis analysis, but tests of heterogeneity across trials can be used to determine which of these need to be modeled as random effects rather than fixed effects. The primary test for heterogeneity of moderation, is shown at the bottom of Column 1. Here we would test for non-zero variance of the trial level moderating effect, Ha: $\text{Var}(\mathbf{b_j}) > 0$. This will determine whether there is any unexplained trial level heterogeneity in comparing the intervention's effect on baseline smokers vs. nonsmokers. No heterogeneity would indicate a consistent variation in effects for smokers versus nonsmokers.

If there is heterogeneity, we have several ways to decompose this effect further. A portion of this heterogeneity may be attributed to measured characteristics of the trial or Z_j, as indicated in the second column of Table 1. To make this model concrete, Z could measure a community's norms

against youth smoking. The coefficient β_1 measures the strength of this trial level covariate in explaining variation in the effect of an individual level moderator X_{ij} (baseline smoking status) across trials. We can also test to see whether trial level variation in moderation by X_{ij} is sufficiently explained by trial level covariate $\mathbf{Z_j}$. In particular, support for this covariate (community norms regarding smoking) explaining trial level moderation is found if we detect little remaining variance across trials, $Var(\mathbf{b_j}| \mathbf{Z_j}) > 0$, once we condition on this trial level covariate.

We can delve further in our assessment of trial level variation in the effects of this moderator. Variation in impact (as a function of smoking status) may also be related to unmeasured characteristics of trials, rather than the measured trial level covariate $\mathbf{Z_j}$ in Column 2. A latent class variable at the trial level is introduced in Column 2, to cluster trials where the moderator effects are similar. Specifically, a set of trial-level latent classes indexed by c are posited, with the proportion of each class given by π_c. One possibility is that trials are conducted is some areas where marijuana use in high school is tolerated and others where it is not, but we have no data on this unmeasured covariate. Then the moderator effect of X_{ij} (baseline smoking) within trial j is given by $\mathbf{b_j}$, which shares a common mean $\theta^{(c)}$ across trials in the same latent class (e.g., in regions where there is low tolerance of marijuana use). It is possible to add additional trial level covariates to this model involving latent classes (Column 4). Here we use a trial level covariate $\mathbf{U_j}$ (say community arrest rate for marijuana use) to predict that trial's class membership, which in turn predicts the trial's moderator effect $\theta^{(c)}$. These models can be fit using two-level models that allow for covariates and discrete mixtures (Asparouhov & Muthen, 2008; Brown, Wang, & Sandler, 2008; Muthen & Asparouhov, 2003a; Muthen & Muthen, 2000; Muthen & Asparouhov, 2008, 2009).

Table 2 provides a similar set of models involving individual level variation (and by extension other levels such as school) that are modeled to hold across all the trials. Note that the variation of moderator effects in Column 1 of Table 2, $\mathbf{b_{ij}}$ (difference in effect of intervention by smoking status) is now treated as random across all individuals and trials; in Table 1 they were treated as constant within trials. Models such as these are best fit when we have repeated outcome measures (e.g. of marijuana use) so that growth modeling can be used to assess changes over time using random slopes and growth mixture modelling (Muthen & Asparouhov, 2003a, 2003b). Another way in which significant heterogeneity can be detected is to investigate how the variance depends on the covariate (Klein & Muthen, 2006), as indicated by the alternative hypothesis at the bottom of Column 1 in Table 2.

Concretely, a finding that there is more variation in growth trajectories of marijuana use among smokers compared to non-smokers who are exposed to intervention ($v > 0$) suggests an unexplained source of variation in impact. The remaining parts of this table follow similarly to that of Table 1. In particular, for the second column of Table 2, we can use a measured individual level covariate (say affiliation with peer smokers) to explain how a particular moderator (baseline smoking) influences this outcome. This corresponds to a three-way interaction between treatment, X, and Z. In the third column, we can model variation in the moderator effect into distinct but unobserved classes. Finally, in the last column these classes are predicted by other measured covariates U. All of these models can be examined using discrete mixtures of random slopes within a latent class framework (Muthen, 2001; Muthen & Asparouhov, 2003b; Muthen & Shedden, 1999; Muthen & Muthen, 2007).

Using Growth Models to Address Different Times of Measurement across Trials

A set of trials will rarely use the exact same times of measurement for outcomes, so one methodologic problem is to calibrate outcomes so they are developmentally comparable. If all the trials use the same instrument to measure response at multiple periods of outcome but the observation times differ across trials, then growth models can be used to standardize the change in response through time. This standardized coding allows intercepts and slopes to have the same

meaning across all trials. Let Y_{ijt} represent the response of subject i in trial j at the tth time point. A linear growth model specifies that these observations are the sum of an individual linear component with random intercept a_{ij} and slope b_{ij}, and a unique independent error ε_{ijt} about this line,

$$Y_{ijt} = a_{ij} + b_{ij}\,\tau_{ijt} + \varepsilon_{ijt}$$

Here τ_{ijt} is the tth time point of observation for this subject. Even though time points may vary across the trials, and across subjects within trials, the modeling of how intercepts and slopes are affected by intervention and other covariates provides a standardized way of assessing change. In particular, a moderation model for the slope that involves covariate X and intervention condition T becomes,

$$b_{ij} = \gamma_0 + \gamma_1 a_{ij} + \gamma_2 X_{ij} + \gamma_3 T_{ij} + \gamma_4 X_{ij} T_{ij} + \varepsilon_{ij}.$$

In this last expression the coefficient γ_4 gives the magnitude of this moderating effect. This model allows the slope to be related to the subject's own intercept, since change is often correlated with one's initial level. A different type of model for moderation is one where the initial level, or intercept, interacts with intervention condition. Here one's change in outcome over time depends on one's latent intercept and intervention.

$$b_{ij} = \delta_0 + \delta_1 a_{ij} + \delta_2 T_{ij} + \delta_3 a_{ij} T_{ij} + \varepsilon_{ij}.$$

The coefficient δ_3 measures the differential effect of intervention on the slope as a function of baseline level. Analytic methods exist for fitting such models. In fact, to obtain good quality fits to the data, we may need to include more terms that allow the random slope to be affected nonlinearly by one's random intercept. The following model provides a quadratic relationship between one's random intercept and slope.

$$b_{ij} = \delta_0 + \delta_1 a_{ij} + \delta_2 T_{ij} + \delta_3 a_{ij} T_{ij} + \delta_4 a_{ij}^2 + \delta_5 a_{ij}^2 T_{ij} + \varepsilon_{ij}$$

By coding treatment status T_{ij} as 1 for intervention and 0 for control, the coefficient δ_4 expresses the quadratic relationship of the intercept on the slope in the control group, and δ_5 as the difference in this quadratic relationship between intervention and control. Thus δ_3 and δ_5 measure the moderating influence of the baseline level on the slope. Computationally these models can be fit using software in Mplus (Klein & Moosbrugger, 2000; Muthen & Asparouhov, 2003b; Muthen & Muthen, 1999).

Three Data Sharing Strategies for Combining Information Across Trials

In the previous section we presented a range of analytic models and methods that could be used in synthesizing moderation findings across trials. The specific modeling that can be done depends on the type of data that are available from the different trials. This availability ranges from published summaries of moderator analyses at one extreme to fully available individual level data from all trials at the other. The quality and precision of the modelling will depend on the level of data and the completeness of data that can be assembled, and this availability depends on the level of data sharing. In this next section we describe three strategies for combining information across trials, based on the degree to which data sharing occurs. We will find that two of these strategies are quite useful in synthesizing findings about moderation.

Standard Meta-Analysis of Moderator Effects with No Sharing of Data

Meta-analysis has been used as a tool for synthesizing study results for more than 30 years (Glass, 1976, 1977; Glass, McGaw, & Smith, 1981; Glass & Smith, 1978; Glass & McAtee, 2006; Smith & Glass, 1977) and is the primary technology used in evidence-based medicine, particularly by the Cochrane Collaboration (Higgens & Green, 2008). In its basic form, meta-analysis combines

published findings across similar studies by placing summary statistics from each of these findings in a common metric, such as an effect size (ES), a standardized mean difference between intervention and control, or relative risk (RR) type measure for dichotomous outcomes. These provide a single measure of impact representing overall effect across trials. Methods for assessing heterogeneity are available (Cook, 1992; Hedges & Olkin, 1985; Wilson & Lipsey, 2003), including those involving random and discrete mixtures (Brown, Wang, & Sandler, 2008).

There has been extensive improvement in the meta-analytic method over the years as it has been extended to examine a range of topics, including the effects of prevention programs on depression (Horowitz & Garber, 2006; Jane-Llopis, Hosman, Jenkins, & Anderson, 2003) and on mental health for youth (Durlak & Wells, 1997), as well as the effect of antidepressants on symptom reduction (Bridge et al., 2007). The method can greatly increase power to detect important effects and reduce the risk to detect effects due to chance. When Bridge and colleagues (Bridge, et al., 2007) began their meta-analysis of antidepressant effects, very few of the 27 existing randomized trials demonstrated significant findings. By combining trials through meta-analysis, Bridge and colleagues demonstrated a highly significant cross-study effect (61% reduction on medication versus 50% on placebo) that was not apparent in single trials.

Systematic review begins with a specification of inclusion and exclusion criteria for trials, followed by an extended search for such trials in the published and fugitive literature, in order to avoid publication bias. Based on information provided in these reports—and occasional clarification with the research team—findings from each trial are combined together by mean of the metanalysis to assess overall effects (Cooper, 1999; Cooper, 2010; Cooper, Hedges, & Valentine, 2009; DerSimonian & Laird, 1986; Hedges & Olkin, 1985; Mosteller & Colditz, 1996). An important strength of this standard meta-analytic review is that great care is taken to identify the complete universe of studies in order to avoid publication bias in estimating an overall effect. Unlike the other two methods described below, there is no requirement for sharing of data for meta-analysis, so main effect analyses for all trials contribute to the assessment of overall intervention effects. If the meta-analysis uses a more limited search, it will preferentially exclude trials with null findings because these are less likely to be published or known.

To conduct a meta-analysis of moderator effects, we would make use of all findings of moderator analyses that are taken from available reports. All those involving the same variable or subgroup would then be combined with standard meta-analytic techniques. Others have raised concerns about the use of meta-analysis to examine moderation (Kraemer et al., 2002; Lipsey, 2003; Shadish & Sweeney, 1991), and we find there are two major limitations with this approach, ones that are so problematic that we do not recommend using this meta-analytic summary strategy for moderation analyses. The first problem is that while all trials can be expected to publish main effect analyses, whether or not they are significant, the same is not true with models involving interactions. It is far more likely for significant interactions to be published and non-significant interactions to be absent from papers, so publication bias in moderator analyses can be large. In a recent meta-analysis on antidepressants, for example, only 9 of 15 trials reported any analyses involving duration of disorder and outcome, so fully 40% of these trials were excluded from this moderation analysis (Bridge, et al., 2007).

The second problem with conducting meta-analyses on moderation analysis is that such analyses can often be conducted quite differently across studies, and therefore these differences in the analytic model make it much more difficult to combine findings. Figure 1a shows that the synthesis step in meta-analysis relies exclusively on input from published summaries that are out of the direct control of the synthesis researchers. This can be especially problematic when examining moderation with continuous covariates, such as baseline risk score. Risk scores can be dichotomized at different

cut-off points; even if they are treated as continuous variables, the need to look for nonlinear effects (Brown, 1993; Brown, Wang, Kellam, et al., 2008; Hastie & Tibshirani, 1990) or transformations of the data lead to different analyses being presented. Furthermore, missing data may be handled differently across studies, and all of these factors contribute to a method variance that would be difficult to account for.

Integrative Data Analysis for Moderator Effects

In contrast to traditional meta-analysis, the integrative data analysis (IDA) strategy assembles all the individual level data into one dataset, treating trial as the highest level in a multilevel modeling framework (see Figure 1b). IDA, which has sometimes been called patient level or individual level meta-analysis, is a type of pooled analysis that has been shown to have great promise in longitudinal research (Bauer & Hussong, 2009; Curran & Hussong, 2009; Curran et al., 2008; Hofer & Piccinin, 2009; Hussong et al., 2008; Shrout, 2009). It has seen limited use so far in randomized trials (Berlin, Santanna, Schmid, Szczech, & Feldman, 2002; Cooper & Patall, 2009; Higgins, Whitehead, Turner, Omar, & Thompson, 2001) and even less in the examination of moderator effects.

Despite great potential, there are three major challenges in this type of integrative data analysis for moderator effects. One challenge involves how to conduct a combined analysis of different datasets all within one analytic model whose outcomes are measured at different times; our discussion of growth models above handles this situation. Another challenging problem occurs when trials use different assessment instruments and covariates. There are procedures that provide some flexible ways of dealing with different measures across trials. These include collapsing of a scale until common categories can be obtained (i.e., similar income categories), using "anchoring items" that are common across different instruments and item response theory to provide comparable scaling (Bauer & Hussong, 2009; Curran & Hussong, 2009; Curran, et al., 2008; Hofer & Piccinin, 2009), and use of missing data procedures.

The third challenge with IDA is procedural; all the relevant datasets must be shared with a research team whose responsibility then involves conducting a full analysis. If some of the datasets are not shared, then an IDA synthesis project can easily introduce selection bias in its findings. Thus IDA typically requires a complete or nearly complete set of trials to contribute their individual level data. In a recent paper that examined different effects of antidepressants as a function of baseline level of depressive symptoms, one synthesis group was only able to obtain individual level data for 6 of 23 trials contacted (Fournier et al., 2010). This suggests high selection bias.

Obtaining individual level data from trial research groups is often hard to accomplish, as there are a number of valid reasons why research groups are often reticent to share their data. These include prior agreements with communities, scientific and ethical committees or participants not to share these data, concerns about confidentiality and misuse of these data, storage of older data in ways that are difficult to access and utilize, incorrect interpretation of the data in the hands of statisticians not involved in the study, prior commitments made by the trialists to permit others to analyze their data, and competition between the data synthesizing group and their own research group over publication of findings. In the past, such concerns have limited access to data and greatly impeded synthesis. Recently, the NIH policy on sharing data (http://grants.nih.gov/grants/policy/data_sharing/) has brought about a cultural shift in collaboration, but there are still enormous challenges.

Parallel Analysis Strategy for Conducting Moderation Analyses

This third strategy for synthesizing data from multiple trials balances the need to conduct equivalent analyses in each of the trials and the very real challenges in full sharing of data. The parallel analysis strategy has each of the respective trial research groups conduct analysis on their own data, following standardized analysis protocols. Results of these analyses done in parallel are then combined into a synthesis, as shown in Figure 1c.

The parallel analysis approach has several advantages and potential challenges. First, because this approach obviates the need for sharing of individual level data and maintains control of the analysis in the hands of the original research group, it is more likely that research groups will be willing to join in the synthesis project. Our experience conducting parallel analyses as part of the United States-European Union (US-EU) Drug Abuse Prevention Project suggests that this approach can improve the likelihood of participation in synthesis projects. This collaborative project funded by the National Institute on Drug Abuse and the European Monitoring Centre for Drugs and Drug Addiction, combined two large drug prevention trials: the US's Adolescent Substance Abuse Prevention Study and the EU's Drug Abuse Prevention Study. It has permitted the examination of moderation effects beyond what could have been accomplished by either of these trials alone.

Just as important, there is an advantage to having the original research team analyze its own data. Because of the trialists' intimate knowledge, they are less likely than those conducting an IDA to misinterpret their own data or conduct flawed analyses. Such occurrences can be commonplace with analysts who are unfamiliar with all the intricacies of the data, and the tacit knowledge available to the original research team is often very difficult to make explicit to outsiders. Some of the challenges in using this method are that the original research team may not have the resources, time, or motivation to conduct these parallel analyses on their own data. Also, some loss of information can come from the combining of the separate analyses in a two-stage analytic procedure, compared to one that combines all the data into one analysis. Also, parallel data analysis makes model checking much more difficult to conduct.

Conclusions

We have noted the very real power limitations in conducting moderation analysis in a single trial, but considerable opportunity to strengthen findings about moderation through combining data from multiple trials. Our conclusion is that unless the heterogeneity across trials is large, the power to detect moderation is almost always increased by combining data across trials compared to that available in a single trial.

In this paper we have laid out a new set of models in Tables 1 and 2 that can be used not only to assess an overall strength of moderation but allow us to examine sources of heterogeneity both within and between trials. These sources may be decomposed into measured as well as unmeasured or unassessed factors that can occur both within and between trials. One important challenge in conducting moderator analyses across multiple trials is calibrating different times of measurement. Growth modeling techniques allow us to summarize growth patterns as random intercepts and slopes whose meaning transcends specific measurement times.

We discussed advantages and disadvantages in using three different ways of combining data across different trials. The traditional meta-analytic method does not use individual level data, and because moderator analyses are often not reported for some trials, or are conducted using different analyses, this method is often not appropriate for synthesis of moderator effects. The other two methods described, integrative data analysis, and parallel analysis, do provide viable choices for synthesis.

This paper has concentrated on analytic modeling, but interpretations of findings have to take account of alternative ways that trials can differ from one another, otherwise there may be little meaning in combining effects. These differences occur in four general categories: individual factors, contextual factors, intervention condition factors, and trial design factors. The most direct to deal with are trial differences in measured individual level factors, i.e., distributions of baseline risk and protective factors. We can account for these through multilevel models in Tables 1 and 2.

Contextual factors, including neighborhood and other socio-cultural factors can have major impact on behavioral outcomes. One interesting example is to assess whether a parent-based training program works equally well when delivered in Spanish to mono-lingual parents, compared to delivery in English (Dillman Carpentier et al., 2007). This question has clear policy implications about whether different intervention components would be needed to deal with the known variations in risk factors for first generation versus later generation immigrants. Analytically, we may test for differences in effectiveness through a test of an interaction; alternatively we could examine whether there is evidence to support similar effects through equivalence testing (Barker, Luman, McCauley, & Chu, 2002). Other contextual differences may be more subtle but relevant. For example, behavioural interventions that targeted HIV risk early in the AIDS epidemic had to overcome more stigma than recent ones, as HIV is now treatable.

Besides these individual and contextual level factors, differences in response across trials can be due to differences in the intervention conditions themselves. Across trials, the intervention conditions can differ in dosage, intensity, fidelity, modality, or person who delivers the intervention. Likewise different trials can vary in their control condition; one school could have no active prevention program while a second may be exposing some of the students to another prevention program.

Finally, differences in the trial designs themselves can lead to variation. We have discussed how to account for different times of measurement and different outcome measures, but sample recruitment and follow-up procedures can also affect findings as well. One important issue for implementation of behavioral interventions is whether a trial is conducted in an efficacy mode, where high fidelity is consistent across the study, or in an effectiveness mode, where larger variations in fidelity can occur.

There are two general ways of handling such differences in intervention conditions across trials. If there is a clear measure that distinguishes interventions, such as duration or dosage, this can be treated as a covariate or moderating factor as shown in Column 2 of Table 1. However, we often do not have sufficient quantitative information to account for these differences, and even if we do, we may still have residual unexplained heterogeneity in moderation. It is always important to allow for and test for trial level variation through multilevel modeling, in both main effect and moderation analyses. The models in Table 1 provide for testing of unexplained heterogeneity in the absence of covariates (Column 1) and in their presence (Column 2). The remaining two columns allow for discrete mixtures, and it may be that both discrete classes and continuous random effects may be needed (Brown, Wang, & Sandler, 2008).

There are a number of limitations to the methods described in this paper. This paper has concerned itself exclusively with the examination of a single moderator variable thought a priori to affect impact. At the other extreme are moderator analysis involving more global subgroup differences in response. One example is in the search for genetic factors that may interact with an intervention. Here we may have upwards of 1,000 candidate markers, coded 0, 1, or 2 depending on the number present in one's DNA, which can be screened for significant effects. One would need to adjust for multiple testing using methods such as false discovery rates (Benjamini & Hochberg, 1995).

If there are a limited number of trials available for understanding moderation, then power to detect heterogeneity in models for Table 1 may be very low. In the United States-European Union (US-EU) Drug Abuse Prevention Project, we have had success, combining data from two randomized trials, provided the trials themselves are large and there are levels of clustering, such as schools within trials, that provide enough degrees of freedom to examine heterogeneity at that level.

The ultimate success or failure of any synthesis project that uses parallel analyses or integrative data analyses hinges on the collaborative partnership that is formed. While there are clear advantages for the science and the public in synthesizing findings across trials, full use of all the data at any given time is often not possible to achieve. Those who have designed these complex studies have commitments to publish results on their studies in a timely fashion, and those related to synthesis projects can either compete for this time or not take into account the unique features necessary to conduct complex modeling that incorporates all the strengths of that particular trial. Handing over data to a centralized analysis unit may lead to incorrect use of these data unless there is an ongoing relationship between the synthesis group and the individual trial groups. On the other hand, synthesis projects can help facilitate the work conducted on the separate trials as well. Such projects can provide additional expertise in methods, and they may uncover unique aspects of one trial relative to others that can then be pursued more effectively through more detailed analyses conducted by that particular research group. This can encourage individual groups to collaborate with the synthesis project, resulting in new research questions and publication opportunities for these research groups. It is also possible to combine statistically the data and findings from all three types of data sharing in one analysis; which may be necessary to accommodate different sharing agreements. All these challenges and opportunities need to be addressed in a synthesis project, so that the partnership fulfills the collective and individual needs.

References

Asparouhov, T.; Muthen, BO. Multilevel mixture models. In: Hancock, GR.; Samuelsen, KM., editors. Advances in latent variable mixture models. Information Age Publishing, Inc; Charlotte, NC: 2008. p. 27-51.

Barker LE, Luman ET, McCauley MM, Chu SY. Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage. American Journal of Epidemiology. 2002; 156:1056–1061. [PubMed: 12446263]

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods. 2009; 14:101–125.[PubMed: 19485624]

Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-(Statistical Methodology). 1995; 57:289–300.

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

Bridge JA, Iyengar S, Salary CB, Barbe RP, Birmaher B, Pincus HA, Brent DA. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: a meta-analysis of randomized controlled trials. Journal of the American Medical Association. 2007; 297:1683–1696. [PubMed: 17440145]

Brown CH. Analyzing preventive trials with generalized additive models. American Journal of Community Psychology. 1993; 21:635–664. [PubMed: 8192125]

Brown CH, Liao J. Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiology paradigm. American Journal of Community Psychology. 1999; 27:673–710. [PubMed: 10676544]

Brown CH, Wang W, Kellam SG, Muthen BO, Petras H, Toyinbo P, The Prevention Science and Methodology Group. Methods for testing theory and evaluating impact in randomized field trials: Intent-to-treat analyses for integrating the perspectives of person, place, and time. Drug and Alcohol Dependence. 2008; 95(Suppl. 1):S74–S104. [PubMed: 18215473]

Brown CH, Wang W, Sandler I. Examining how context changes intervention impact: The use of effect sizes in multilevel meta-analysis. Child Development Perspectives. 2008; 2:198–205. [PubMed: 20585469]

Brown EC, Hawkins JD, Arthur MW, Briney JS, Abbott RD. Effects of Communities That Care on prevention services systems: findings from the community youth development study at 1.5 years. Prevention Science. 2007; 8:180–191. [PubMed: 17602298]

Coie JD, Watt NF, West SG, Hawkins JD. The science of prevention: A conceptual framework and some directions for a national research program. American Psychologist. 1993; 48:1013–1022. [PubMed: 8256874]

Cook, TD. Meta-analysis for explanation : A casebook. Russell Sage Foundation; New York: 1992. Cooper, H. Synthesizing research: A guide for literature reviews. 3rd ed.. SAGE Publications; Thousand Oaks, CA: 1999.

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

Cooper, HM. Research synthesis and meta-analysis: A step-by-step approach. 4th ed.. SAGE Publications; Los Angeles, CA: 2010.

Cooper, HM.; Hedges, LV.; Valentine, JC. The handbook of research synthesis and meta-analysis. 2nd ed.. Russell Sage Foundation; New York: 2009.

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. [PubMed: 19485623]

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. Developmental Psychology. 2008; 44:365–380. [PubMed: 18331129]

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7:177–188. [PubMed: 3802833]

Dillman Carpentier FR, Mauricio AM, Gonzales NA, Millsap RE, Meza CM, Dumka LE, Genalo MT. Engaging Mexican origin families in a school-based preventive intervention. Journal of Primary Prevention. 2007; 28:521–546. [PubMed: 18004659]

Dishion, TJ.; Burraston, B.; Poulin, F. Peer group dynamics associated with iatrogenic effects in group interventions with high-risk young adolescents. In: Erdley, C.; Nangle, DW., editors. Damon's new directions in child development: The role of friendship in psychological adjustment. Jossey-Bass; San Francisco: 2001. p. 79-92.

Dishion TJ, McCord J, Poulin F. When interventions harm. Peer groups and problem behavior. American Psychologist. 1999; 54:755–764. [PubMed: 10510665]

Dishion TJ, Spracklen KM, Andrews DW, Patterson GR. Deviancy training in male adolescents friendships. Behavior Therapy. 1996; 27:373–390.

Dolan LJ, Kellam SG, Brown CH, Werthamer-Larsson L, Rebok GW, Mayer LS, Wheeler L. The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. Journal of Applied Developmental Psychology. 1993; 14:317–345.

Durlak JA, Wells AM. Primary prevention mental health programs for children and adolescents: A meta-analytic review. American Journal of Community Psychology. 1997; 25:115–152. [PubMed: 9226860]

Elliott DS, Mihalic S, Elliott DS, Mihalic S. Issues in disseminating and replicating effective prevention programs. Prevention Science. 2004; 5:47–53. [PubMed: 15058912]

Ennett ST, Tobler NS, Ringwalt CL, Flewelling RL. How effective is drug abuse resistance education? A meta-analysis of Project DARE outcome evaluations. American Journal of Public Health. 1994; 84:1394–1401. [PubMed: 8092361]

Faggiano F, Vigna-Taglianti F, Versino E, Zambon A, Borraccino A, Lemma P. School-based prevention for illicit drugs' use. Cochrane Database of Systematic Reviews. 2005; 2 Art. No.: CD003020.

Faggiano F, Vigna-Taglianti FD, Versino E, Zambon A, Borraccino A, Lemma P. School-based prevention for illicit drugs use: a systematic review. Preventive Medicine. 2008; 46:385–396. [PubMed: 18258289]

Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, Ji P. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prevention Science. 2005; 6:151–175. [PubMed: 16365954]

Fournier JC, DeRubeis RJ, Hollen SD, Dimdjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant Drug Effects and Depression Severity: Patient-level Meta-Analysis. Journal of the American Medical Association. 2010; 303:47–53. [PubMed: 20051569]

Glass GV. Primary, secondary, and meta-analysis of research. Educational Researcher. 1976; 5(10):3–8.

Glass GV. Integrating findings: The meta-analysis of research. Review of Research in Education. 1977; 5:351–379.

Glass, GV.; McGaw, B.; Smith, ML. Meta-analysis in social research. SAGE Publications; Beverly Hills: 1981.

Glass, GV.; Smith, ML. Meta-analysis of research on the relationship of class-size and achievement. The class size and instruction project. National Inst. of Education (DHEW); Washington, DC: 1978. p. 81

Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. Social Science & Medicine. 2006; 62:1650–1671. [PubMed: 16198467]

Hastie, T.; Tibshirani, R. Generalized additive models. 1st ed.. Chapman and Hall; London, New York: 1990.

Hedges, LV.; Olkin, I. Statistical methods for meta-analysis. Academic Press; Orlando, FL: 1985.

Higgins, JPT.; Green, S., editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration; 2008. Retrieved from http://www.cochrane-handbook.org

Higgins JP, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Statistics in Medicine. 2001; 20:2219–2241. [PubMed: 11468761]

Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychological Methods. 2009; 14:150–164. [PubMed: 19485626]

Horowitz JL, Garber J. The prevention of depressive symptoms in children and adolescents: A metaanalytic review. Journal of Consulting and Clinical Psychology. 2006; 74:401–415. [PubMed: 16822098]

Howe GW, Reiss D, Yuh J. Can prevention trials test theories of etiology? Development and Psychopathology. 2002; 14:673–693. [PubMed: 12549699]

Hussong A, Cai L, Curran P, Flora D, Chassin L, Zucker R. Disaggregating the distal, proximal, and time-varying effects of parent alcoholism on children's internalizing symptoms. Journal of Abnormal Child Psychology. 2008; 36:335–346. [PubMed: 17891557]

Ialongo N, Poduska J, Werthamer L, Kellam S. The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. Journal of Emotional & Behavioral Disorders. 2001; 9:146–160.

Ialongo NS, Werthamer L, Kellam SG, Brown CH, Wang S, Lin Y. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. American Journal of Community Psychology. 1999; 27:599–641. [PubMed: 10676542]

Jane-Llopis E, Hosman C, Jenkins R, Anderson P. Predictors of efficacy in depression prevention programmes. Meta-analysis. British Journal of Psychiatry. 2003; 183:384–397. [PubMed: 14594912]

Kellam SG, Brown CH, Poduska JM, Ialongo NS, Wang W, Toyinbo P, Wilcox HC. Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. Drug and Alcohol Dependence. 2008; 95(Suppl.1):S5–S28. [PubMed: 18343607]

Kellam SG, Koretz D, Moscicki EK. Core elements of developmental epidemiologically based prevention research. American Journal of Community Psychology. 1999; 27:463–482. [PubMed: 10573831]

Kellam SG, Langevin DJ. A framework for understanding "evidence" in prevention research and programs. Prevention Science. 2003; 4:137–153. [PubMed: 12940466]

Klein A, Moosbrugger H. Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika. 2000; 65:457–474.

Klein AG, Muthen B. Modeling heterogeneity of growth depending on initial status. Journal of Educational and Behavioral Statistics. 2006; 31:357–375.

Kraemer HC, Wilson GT, Fairburn CG, Agras WS, Kraemer HC, Wilson GT, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry. 2002; 59:877–883. [PubMed: 12365874]

Lipsey MW. Those confounded moderators in meta-analysis: Good, bad, and ugly. Annals of the American Academy of Political and Social Science. 2003; 587:69–81.

Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). Annual Review of Public Health. 1996; 17:1–23.

Muthen, B. Latent variable mixture modeling. In: Marcoulides, GA.; Schumacker, RE., editors. New developments and techniques in structural equation modeling. Lawrence Erlbaum Associates; Hillsdale, NJ: 2001. p. 1-33.

Muthen B, Asparouhov T. Advances in latent variable modeling, Part I: Integrating multilevel and structural equation modeling using Mplus. 2003a Manuscript in Preparation.

Muthen B, Asparouhov T. Modeling interactions between latent and observed continuous variables using maximum-likelihood estimation in Mplus. Mplus Web Notes, #6. 2003b Retrieved from http://www.statmodel.com/download/webnotes/webnote6.pdf.

Muthen B, Muthen LK. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. Alcoholism: Clinical & Experimental Research. 2000; 24:882–891.

Muthen, BO.; Asparouhov, T. Growth mixture modeling: Analysis with non-Gaussian random effects. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. Longitudinal data analysis: A handbook of modern statistical methods. Boca Raton, FL; Chapman & Hall/CRC Press; 2008. p. 143-165.

Muthen BO, Asparouhov T. Multilevel regression mixture analysis. Journal of the Royal Statistical Society. 2009; 172:639–657.

Muthen BO, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999; 55:463–469. [PubMed: 11318201]

Muthen, LK.; Muthen, BO. Mplus user's guide version 2. 2nd ed.. Muthen and Muthen; Los Angeles, CA: 1999.

Muthen, LK.; Muthen, BO. Mplus: Statistical analysis with latent variables: User's guide. 4th ed.. Muthen & Muthen; Los Angeles, CA: 2007.

O'Connell, ME.; Boat, T.; Warner, KE., editors. Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities. National Academy Press; Washington, DC: 2009.

Shadish WR, Sweeney RB. Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. Journal of Consulting and Clinical Psychology. 1991; 59:883–893. [PubMed: 1837794]

Shrout PE. Short and long views of integrative data analysis: Comments on contributions to the special issue. Psychological Methods. 2009; 14:177–181. [PubMed: 19485628]

Sloboda Z, Stephens RC, Stephens PC, Grey SF, Teasdale B, Hawthorne RD, Marquette JF. The Adolescent Substance Abuse Prevention Study: A randomized field trial of a universal substance abuse prevention program. Drug and Alcohol Dependence. 2009; 102:1–10. [PubMed: 19332365]

Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. American Psychologist. 1977; 32:752–760. [PubMed: 921048]

Tein J-Y, Sandler IN, MacKinnon DP, Wolchik SA. How did it work? Who did it work for? Mediation in the context of a moderated prevention effect for children of divorce. Journal of Consulting & Clinical Psychology. 2004; 72:617–624. [PubMed: 15301646]

Tobler NS. Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcome results of program participants compared to a control or comparison group. Journal of Drug Issues. 1986; 16(4):537–567.

Van Horn ML, Jaki T, Masyn K, Ramey SL, Smith JA, Antaramian S. Detecting differential effects: Applying regression mixture models to identify variations in the influence of family resources on academic achievement. Developmental Psychology. 2009; 45:1298–313. [PubMed: 19702393]

Wilson, DB.; Lipsey, MW. The role of method in treatment effectiveness research: Evidence from meta-analysis. In: Kazdin, AE., editor. Methodological issues & strategies in clinical research. 3rd ed.. American Psychological Association; Washington, DC: 2003. p. 589-615.

Wolchik, S.; Sandler, I.; Weiss, L.; Winslow, E. New Beginnings: An empirically-based intervention program for divorced mothers to help children adjust to divorce. In: Briesmeister, JM.; Schaefer, CE., editors. Handbook of parent training : Helping parents prevent and solve problem behaviors. 3rd ed.. John Wiley & Sons; Hoboken, N.J.: 2007. p. 25-66.
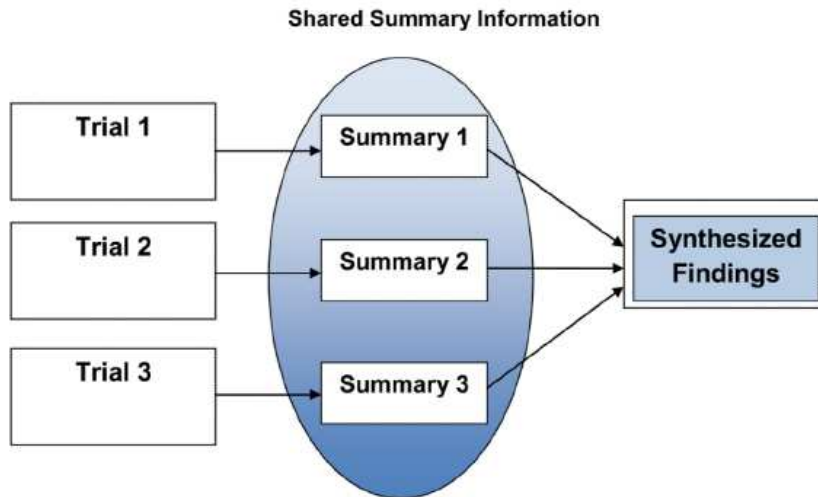
Figure 1a.
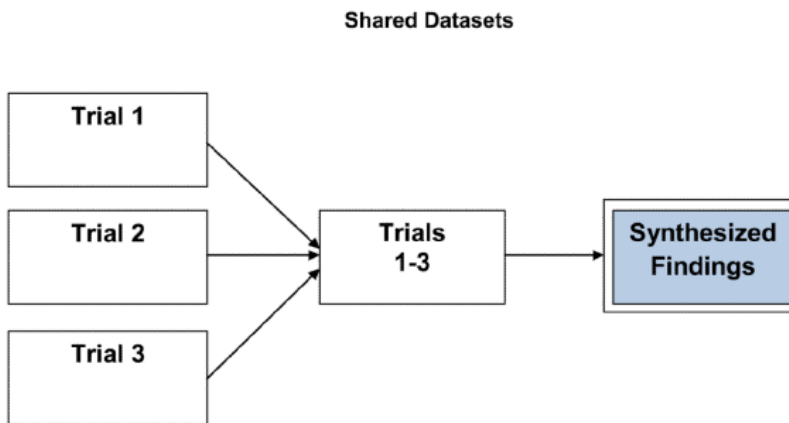The Process of Synthesis using Traditional Meta-Analysis

Figure 1b.
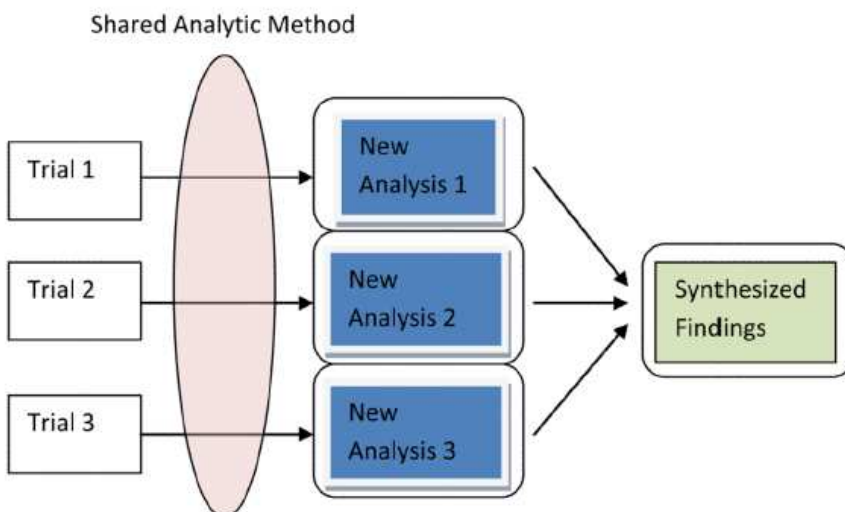Schematic of Synthesis in Integrative Data Analysis.

Figure 1c.
Parallel Analysis Strategy for Synthesizing Moderation Analyses.

Table 1

Analytic Models Examining Variation in Moderation Effects across Trials

Total Heterogeneity across Trials	Measured Trial Level Covariates	Related to Trial Level Latent Classes	Predictors of Trial Level Classes
Y_{ij} $= a_j + a_j^0 X_{ij} + a_j^1 T_{ij}$ $+ b_j X_{ij} T_{ij} + e_{ij}$	$b_j = \beta_0 + \beta_1 Z_j + \varepsilon_j$	$\Pr(C_j = c) = \pi_c$ $b_j = \theta + \theta^{(c)} + \varepsilon_i$	logit $\Pr(C_j = c) =$ $\lambda_0^{(c)} + \lambda_1^{(c)} U_j$ $b_j = \theta + \theta^{(c)} + \varepsilon_i$
H_a: $Var(b_j) > 0$	H_a: $\beta_1 \neq 0$ H_a: $Var(b_j \mid Z_j) > 0$	H_a: $\theta^{(c)} \neq 0$	H_a: $\lambda_1^{(c)} \neq 0$

Table 2

Analytic Models Examining Variation in Moderation Effects Involving Individual Level Factors Consistent Across Trials

Total Heterogeneity of Moderation across Individuals and Trials	Measured Individual Level Covariates	Related to Individual Level Latent Classes	Predictors of Individual Level Classes
Y_{ij} $= a + a_{ij}^0 X_{ij} + a_{ij}^1 T_{ij}$ $+ b_{ij} X_{ij} T_{ij} + e_{ij}$	$b_{ij} = \beta_0 + \beta_1 Z_{ij} + e_j$	$\Pr(C_{ij} = c) = \pi_c$ $b_{ij} = \theta + \theta^{(c)} + e_{ij}$	logit $\Pr(C_{ij} = c)$ $= \lambda_0^{(c)} + \lambda_1^{(c)} U_{ij}$
H_a: $Var(b_{ij} \mid X_{ij}) =$ $u + v X_{ij}^2$, $v > 0$	H_a: $\beta_1 \neq 0$	H_a: $\theta^{(c)} \neq 0$	H_a: $\lambda_1^{(c)} \neq 0$