AN APPLICATION OF FINITE MIXTURE MODELING TO CHARACTERIZE SOURCES OF BETWEEN-STUDY VARIATION IN META-ANALYSES OF PREVENTION PROGRAM EFFECTS

by

NICHOLAS J. PARR

A DISSERTATION

Presented to the Department of Counseling Psychology and Human Services
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Nicholas J. Parr

Title: An Application of Finite Mixture Modeling to Characterize Sources of Between-Study Variation in Meta-Analyses of Prevention Program Effects

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Counseling Psychology and Human Services by:

John R. Seeley	Chair
Emily E. Tanner-Smith	Core Member
Katherine E. Masyn	Core Member
Kathleen Scalise	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020.

DISSERTATION ABSTRACT

Nicholas J. Parr

Doctor of Philosophy

Department of Counseling Psychology and Human Services

June 2020

Title: An Application of Finite Mixture Modeling to Characterize Sources of Between-Study Variation in Meta-Analyses of Prevention Program Effects

In meta-analyses of prevention programs, findings of primary research studies are pooled to estimate an overall program effect, an approach generally offering improved statistical power and precision over analyses at the individual study level. Across studies, however, programs are often implemented with considerable variation in implementation quality, program components, assessment approaches, and sample characteristics. Differences across these and other aspects of a program's implementation can induce between-study variation in program effects. Excessive between-study variation can compromise the utility of a summary estimate of program effect, as derived in meta-analysis, because the estimate can be unrepresentative of the broad distribution of effects across implementations of the program. Importantly, variation produced by observable primary study characteristics is often explainable using variables that represent study methodology, program design, and sample attributes, and utilizing approaches such as subgroup analysis and meta-regression can provide insight into study-level factors that moderate the magnitude or direction of program effects. While widely used, these moderation analysis methods have recognized statistical and interpretive limitations, in particular when there is an interest in understanding the interrelation of multiple potential moderator variables and their combined influence on variation in program effects, as well

as their co-occurrence in typical studies implementing a program. To address these limitations, this dissertation describes and demonstrates a multivariate approach to moderation analysis in aggregate-data meta-analysis, which employs finite mixture modeling as its underlying analytic framework. Results of the approach suggest it provides insight into the co-occurrence of potential moderators in a sample of studies implementing a prevention program, and into how such co-occurrence relates to program effectiveness.

CURRICULUM VITAE

Name of author: Nicholas J. Parr

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Tulane University School of Public Health and Tropical Medicine, New Orleans
Tulane University, New Orleans

DEGREES AWARDED:

Doctor of Philosophy in Prevention Science, 2020, University of Oregon
Master of Science in Prevention Science, 2019, University of Oregon
Master of Public Health in Health Systems and Development, 2013, Tulane University School of Public Health and Tropical Medicine
Bachelor of Arts in Art History, 2008, Tulane University

AREAS OF SPECIAL INTEREST:

Sexual and Gender Minority Health and Wellbeing
Methodologies for Research Synthesis and Effective Program Implementation
Etiology and Prevention of Suicidality

PROFESSIONAL EXPERIENCE:

Senior Program Analyst (HIV, STI, and Viral Hepatitis), National Association of County and City Health Officials, Washington, DC, 2016-2017

Prevention Resources Coordinator, CrescentCare Health and Wellness, New Orleans, 2013-2015

Counseling, Testing, and Outreach Specialist, CrescentCare Health and Wellness, New Orleans, 2010-2013

National HIV Behavioral Surveillance Data Collector, Centers for Disease Control and Prevention, 2012

GRANTS, AWARDS, AND HONORS:

Garrett Lee Smith Suicide Prevention Grant, Substance Abuse and Mental Health Services Administration, Department of Health and Human Services (DHHS), 2018-present (co-investigator)

Faculty-Appointed Graduate Student Representative, Social Systems Data Science Network, University of Oregon, 2020

Faculty-Appointed Representative, College of Education Dean's Student Advisory Board, University of Oregon, 2017-2019

First Year Fellowship, University of Oregon, 2017-2018

Ryan White HIV/AIDS Program Building Care and Prevention Capacity: Addressing the HIV Care Continuum in Southern Metropolitan Areas, Health Resources and Services Administration, DHHS, 2016-2017 (project manager)

Cum Laude, Tulane University, 2008

PUBLICATIONS:

Parr, N. J. (2020). Sexual assault and co-occurrence of mental health outcomes among cisgender female, cisgender male, and gender minority U.S. college students. *Journal of Adolescent Health*. Advance online publication.

Parr, N. J. & Howe, B. G. (2020). Factors associated with frequency of gender identity nonaffirmation microaggressions among transgender persons. *Culture, Health, and Sexuality*. Advance online publication.

Parr, N. J., Schweer-Collins, M. L., Darlington, T. M., and Tanner-Smith, E. E. (2019). Meta-analytic approaches for examining complexity and heterogeneity in studies of adolescent development. *Journal of Adolescence*, *77*, 168-178.

Parr, N. J. & Howe, B. G. (2019). Heterogeneity of transgender identity nonaffirmation microaggressions and their association with depression symptoms and suicidality among transgender persons. *Psychology of Sexual Orientation and Gender Diversity*, *6*(4), 461–474.

Parr, N. J. (under first review). Differences in the age-varying association of school belonging with flourishing among minority and non-minority university students. *Journal of American College Health*.

Parr, N. J. & Howe, B. G. (under first review). Stigmatizing microaggressions, depression symptoms, and suicide ideation among a geographically-diverse sample of sexual minority persons. *Journal of Gay & Lesbian Mental Health.*

ACKNOWLEDGMENTS

I extend my profound gratitude to Dr. John Seeley for his unflagging support and guidance from the earliest moment of my interest in the Prevention Science doctoral program through to its conclusion. I express great appreciation for the cultivation of my research interests, quantitative skillsets, and professional acumen by Drs. Emily Tanner-Smith and Kathleen Scalise. I am also grateful to Dr. Tanner-Smith for access to the data used to demonstrate the proposed methodology. Finally, I offer my thanks to Dr. Katherine Masyn, who despite considerable geographic distance, was a ready voice of clarity and encouragement throughout the development of the present dissertation.

I also extend my appreciation to Dr. Jonathan Minton of the National Health Service, Scotland, who served as a valuable sounding board in the earliest conceptualizing of this project.

Finally, I acknowledge the unwavering companionship of my husband, Brent Pafford, without whose presence and sacrifices I could not have accomplished this endeavor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

In meta-analyses of prevention interventions, policies, or programs (hereafter "programs"), findings of primary research studies are pooled to estimate an overall program effect. Synthesizing effects drawn from numerous primary studies can offer improved statistical power to detect significant effects and provide more precise effect estimates compared with analyses in individual primary studies (Borenstein et al., 2009). At the same time, the implementation of a program in each study can vary in rigor (e.g., differing levels of fidelity monitoring or quality of randomization), setting or sample characteristics (e.g., implementation in controlled clinical vs. naturalistic community site), outcome assessment (e.g., differences in assessment measures or length of follow up), program design (e.g., variation in active components or delivery modality), or in other implementation attributes, and these differences can induce between-study variation in program effects (Higgins & Thompson, 2002). Excessive between-study variation can compromise the utility of a summary estimate of program effect, as derived in meta-analysis, because the estimate can be unrepresentative of the broad distribution of effects across implementations of the program.

Between-study variation in effects produced by observable primary study characteristics is often explainable (Parr et al., 2019; Viechtbauer, 2007). That is, variables representing study attributes may be assessed for their relation with the overall or summary effect estimate, and those found to explain substantial between-study variation can be utilized as statistical controls in meta-analytic models as well as substantively interpreted to provide insight into study-level factors that influence the

magnitude or direction of program effects (Baker et al., 2009; Parr et al., 2019). Several methods have been developed to assess the relation of these study attribute variables, known as moderators, with summary effect estimates.

Traditional methods of moderator analysis include *meta-regression*, which assesses the linear relation of one or more moderators with an average program effect, and *subgroup analysis*, which extends analysis of variance methods to compare effect size estimates calculated among subgroups of studies defined by an observed categorical moderator (e.g., randomized controlled vs. quasi-controlled studies) (Baker et al., 2009; Borenstein & Higgins, 2013; Thompson & Higgins, 2002). When there is an interest in understanding the influence of program implementation rigor, for instance, meta-regression can be used to assess the moderation effects of level of interventionist training, degree of program fidelity monitoring, and availability of implementation support. The outputs of such a model would include coefficients indicating the degree to which the magnitude of the summary effect estimate is altered (moderated) by the effect of each implementation attribute while controlling for other attributes. Alternatively, subgroup analysis could be used to group studies by a relevant categorical moderator, such as presence or absence of fidelity monitoring, and differences in the summary program effect as estimated among studies in each subgroup could be compared.

Both meta-regression and subgroup analysis have several limitations. Subgroup analysis, for example, is limited to exploring effect moderation by a single variable (i.e., the grouping variable). Meta-regression, by contrast, is theoretically able to accommodate unlimited moderators, but as with linear regression more generally, in practice increasing the number of potential moderator variables can lead to multicollinearity and the

possibility of biased and misleading results (Berlin & Antman, 1992). Moreover, with meta-regression it can be necessary to examine whether moderator variables would be better entered into the model in higher-order (polynomial) forms, and whether interactions among moderators should be included (Viechtbauer, 2007). The resulting complexity of an extended meta-regression model, despite the possibility of it being well-fitting and highly explanatory, can hamper its interpretive value to implementation researchers and practitioners.

Beyond the above modeling and interpretation issues (and others to be discussed in following chapters), in the application of investigating moderators of program effectiveness, meta-regression and subgroup analysis are limited in their capability to answer a central question of program implementation research: What are the general set of characteristics of a program's implementation, that when applied together, enhance the program's effectiveness? Continuing the above example, the applied researcher or practitioner may find use in understanding the discrete effect of fidelity monitoring while holding constant the effect of other implementation factors, but considerably greater utility might be found in understanding the impact of a program implemented in a broadly rigorous fashion: utilizing intensive interventionist training, providing program fidelity monitoring, *and* offering extensive implementation support. Investigating whether such factors are influential, on the whole, requires a multivariate analytic approach that can more straightforwardly characterize program implementation across a number of dimensions over which it can typically vary.

Whether implementation characteristics are related to (moderate) a program's effect is a distinct question from how those characteristics typically occur among

implementations of a program across different primary research studies or controlled evaluations. The former question can be investigated with methods such as meta-regression or subgroup analysis, which assess the magnitude and significance of relations between moderators and program effects. At the same time, however, these methods offer little insight into how moderators relate *to one another*, and whether that interrelationship differentiates high- and low-impact program implementations. By contrast, a multivariate method that estimates the co-occurrence of multiple moderators representing program implementation characteristics – and then allows for assessing the relation of that co-occurrence with program effects – would facilitate development of profiles representing the interrelation of multiple moderators and provide greater understanding of their combined influence on program effectiveness.

Organization of the Dissertation

Taken together, the above considerations suggest the need for a multivariate framework for investigating effect size moderation in meta-analysis. The following dissertation describes such an approach employing finite mixture modeling.[1] The dissertation is organized as follows. First, between-study variation in program effects is defined both statistically and as a parameter of substantive interest in meta-analysis. Approaches for quantifying and assessing the impact of this variation are next described, followed by further discussion of existing methods of explaining between-study variation, in particular meta-regression and subgroup analysis. Strengths and limitations

[1] Because of an interest in the applicability of the proposed method to the most common meta-analytic settings, the present dissertation employs aggregate rather than individual participant data. Moreover, the proposed method utilizes a frequentist framework for meta-analysis; potential extensions of the method using prior information in a Bayesian estimation approach are discussed in Chapter V.

of these methods are highlighted. Descriptions of the motivating example, brief substance use interventions, and associated data are then provided. After this overview, the proposed method is considered in detail, beginning with finite mixture modeling in the form of latent class analysis, which serves as the underlying analytic framework for the approach. Additional aspects of the method, including the use of a structural equation modeling framework for estimating meta-analytic models and the procedure for fitting and evaluating final models, are then described. The method is next demonstrated using data derived from a systematic review of studies examining brief substance use intervention effectiveness. The dissertation concludes with a discussion of research and practice implications of the method and its results, as well as important limitations of the approach and future research directions.

Note on terminology: In the following sections, the term heterogeneity, which in meta-analysis typically refers to between-study variation in program effects, is generally avoided given its varying definitions in meta-analytic, latent class and finite mixture, and structural equation modeling literatures, all of which are called upon in the present dissertation. For clarity, therefore, the term *between-study variation* is used to refer to the type of variation in program effects the proposed method is intended to characterize; heterogeneity is reserved for discussion of distributional variation in the finite mixture modeling context.

CHAPTER II

BACKGROUND

Effect Size Variation in Aggregate Data Meta-Analysis

Principally, two forms of variation are present in meta-analysis: within-study error and between-study variation in effects.[2] In *fixed-effect* meta-analysis, a program is assumed to have a single, *common* true effect; deviations from the common true effect found between individual primary studies are assumed to arise from random sources, such as measurement or sampling error within each study (Borenstein et al., 2010). In *random-effects* meta-analysis, by contrast, a program's true effect is allowed to vary across primary studies; differences in the magnitude or direction of effects between each primary study are assumed to result from both within-study error and true differences in the program effect between studies. As a result, in random-effects meta-analysis, study-level effects are conceptualized as drawn from a population distribution of effects, with the studies present in a meta-analytic dataset representing a sample of effects from that population.[3] This distinction is evident in the formulations of the respective effect sizes. For the fixed-effect model, let Y_i be the observed effect size for each study,

$$Y_i = \theta + \epsilon_i \ , \tag{1}$$

[2] *Effect sizes* refers to estimates of effects synthesized in a meta-analysis. These estimates are not limited to measures of program effectiveness, and may be defined as, among other values, prevalence estimates or incidence rates, estimates of diagnostic test sensitivity or specificity, or correlation or regression coefficients.

[3] This is the case when each primary study contributes a single, independent estimate of a program's effect. Primary studies may contribute several, dependent effect sizes that represent multiple measurements of a program's effect (e.g., different measures of alcohol consumption for a program intended to reduce alcohol misuse, and/or effect sizes from multiple time points). For simplicity of exposition, here each study (k) is assumed to contribute one independent effect size, and therefore $k = i$.

which is composed of the common true effect θ and within-study error ϵ_i. In the fixed-effect model, ϵ_i represents the difference between the study's observed effect size and the true common effect θ. In the random-effects model,

$$Y_i = \mu + \xi_i + \epsilon_i \ , \tag{2}$$

the single common effect θ is replaced by $\mu + \xi_i$, where μ is the mean of the (population) distribution of effects and ξ_i is the extent to which the individual study's true effect departs from the population average effect μ. Although the error term ϵ_i again reflects within-study error, in contrast to the fixed-effect model this quantity now represents the difference between each study's observed effect size and study-specific true effect θ_i ($\mu + \xi_i$). In both models, random error ϵ_i is typically assumed to be normally distributed, with mean of zero and variance equal to the observed sampling variance of the effect size, which is treated as known. In the random-effects model, the effect distribution from which each study's ξ parameter is sampled (i.e., $\xi_1, \ldots, \xi_k$) is also assumed normally distributed and the variance of this distribution is τ^2, the magnitude of between-study variation in effects (Hedges & Vevea, 1998). Moreover, each study's ϵ and ξ parameter is assumed independently and identically distributed (Viechtbauer, 2005).

In both fixed- and random-effects meta-analysis, the summary effect (the sample estimate of θ or μ, respectively) is typically calculated simply as a weighted mean of the observed study-level effects (Hedges & Vevea, 1998). The weights employed are usually defined as the inverse of the variance of each effect size. In fixed-effect meta-analysis, the variance is composed of only within-study sources of random error. In the random-effects model, by contrast, between-study variation also comprises the effect size

variance. Contrasting the two weighting formulations, in which W_i is the weight given to each effect size i, the fixed-effect weights are defined as

$$W_i = \frac{1}{V_i} \ , \tag{3}$$

with V_i corresponding to the sample estimate of the within-study error variance v_i. Weights under the random-effects model are calculated as

$$W_i = \frac{1}{V_i + T^2} \ , \tag{4}$$

with V_i defined as in (3) and T^2 defined as the sample estimate of the magnitude of between-study variation in effects (τ^2). Note that V_i is specific to each effect size, whereas T^2 is treated as a constant once estimated. For either model, then, the summary effect $\bar{Y}$ can be calculated as a weighted mean with weights defined as in (3) or (4):

$$\bar{Y} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \ . \tag{5}$$

The differing weighting schemes underline two important distinctions in how the respective models view the effect data under consideration. First, because the random-effects model conceptualizes the data as drawn from a population of effect sizes and incorporates characteristics of that population distribution (specifically, an estimate of its first moment in the form of the summary estimate and its second moment via weighting that incorporates the between-study variance component), random-effects meta-analysis permits *unconditional* inference, or generalization of findings beyond the studies sampled in a meta-analytic dataset (Borenstein et al., 2010; Hedges & Vevea, 1998). Such inference is possible because the estimates of the population parameters allow one to

assess how representative the effect size data is of the distribution of effect sizes from which they were sampled. The fixed-effect model, on the other hand, does not include estimates of a population parameter distribution (because it explicitly assumes there is no distribution), and is therefore confined to *conditional* inference – inference limited to only the studies included in the analysis (Hedges & Vevea, 1998).[4]

The second important distinction is the more salient for the proposed methodology. In fixed-effect meta-analysis, variation in the effect data is essentially uninformative; put another way, because this variation is seen to arise mainly from error in measurement or sampling, the source of between-study variation in effect size estimates is, by definition, explained (as random error). In contrast, by acknowledging (and explicitly estimating) between-study variation in effect sizes, random-effects meta-analysis allows that such variation may arise from sources other than error (Viechtbauer, 2007). As a result, under the random-effects model, between-study variation becomes a potentially informative parameter when sources of that variation are observable.

Quantifying and Explaining Between-Study Variation

To assess whether substantial between-study variation exists in a sample of effect sizes, the τ^2 parameter must first be estimated using either a method of moments or iterative estimation approach. Among the former, the most common is the DerSimonian and Laird (1986) method; among the latter, maximum-likelihood estimation (MLE) and

[4] Hedges and Vevea (1998) provide an exception to this statement, namely when studies outside the sample are identical to studies included in the sample. Given that such a scenario is generally unrealistic, this rationale has been approximated by generalizing to studies determined a priori to be sufficiently similar to the included studies. Hedges and Vevea (1998) note that while the former, identical-study case is well within technical sampling theory for fixed-effect modeling, the latter approach is extrastatistical and vulnerable to bias.

restricted maximum-likelihood (REML) estimation are frequently employed (Cheung, 2015; Viechtbauer, 2005). Each method has strengths and limitations with regard to efficiency and downward biasedness in the estimate of τ^2, with the REML estimator found to best balance both considerations (i.e., maximizing efficiency and minimizing downward biasedness). Unrestricted MLE can underestimate τ^2 when the total number of studies is small (Viechtbauer, 2005), but MLE is sometimes preferred for its versatility and compatibility with other modeling approaches. When random-effects meta-analysis is carried out in a structural equation modeling (SEM) framework (see Chapter III), for instance, the within-study true effect θ_i is treated as a latent random variable whose variance (τ^2) is estimated using MLE (Cheung, 2015). Further discussion regarding the use of MLE (and related concerns about the possible underestimation of τ^2) in the context of the proposed methodology is provided in Chapter III, below.

When the estimate of τ^2 indicates between-study variation in effect sizes, several strategies can be employed to identify potential sources of between-study variation. Perhaps the most straightforward method is to conceive of the effect sizes as belonging to one of two or more groups defined by the levels of a categorical moderator variable. Once effect sizes are grouped accordingly, the summary effect can be calculated in each group using a typical fixed- or random-effects model and compared. This method is referred to as subgroup analysis (e.g., Borenstein & Higgins, 2013), and in the random-effects application study effects sizes Y_i are estimated as

$$Y_i|G = \mu_g + \xi_{ig} + \epsilon_{ig} \ , \tag{6}$$

with G denoting the observed subgroup of primary studies defined by the categorical moderator variable, and μ_g, ξ_{ig}, and ϵ_{ig} representing the same quantities defined as in (2) but only among the studies classified within each subgroup. Subgroup analysis is often employed in meta-analyses of clinical trial data; with such data, there is typically interest in a single moderator, for instance levels of a pharmaceutical treatment. Subgroup analysis in this context might involve subgroups based on whether studies provided, for example, a 500 mg, 1000 mg, or 1500 mg dosage of a medication. The summary treatment effect and between-study variance would then be estimated in each subgroup,[5] and differences in the summary effect between subgroups assessed for statistical significance using adaptations of analysis of variance or t-tests (Borenstein et al., 2009).

In the social sciences, there are frequently several (even many) potential moderators of interest, and traditional subgroup analysis becomes limiting. These moderators may be selected based on analyst expertise, a priori hypotheses, previous findings, or in some cases, using a post hoc selection process (Baker et al., 2009; Parr et al., 2019).[6] An alternative to subgroup analysis compatible with multiple moderators is meta-regression, which for consistency with fixed- and random-effects model terminology, may also be described as a mixed-effects model for meta-analysis. For each study's effect size, the meta-regression model generally takes the form of

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \xi_i + \epsilon_i \ , \tag{7}$$

[5] If between-study differences in effect are assumed to be equivalently distributed in all subgroups, a common estimate of between-study variation (τ^2) across groups can be estimated (Borenstein et al., 2009).

[6] A growing area of interest is algorithmic moderator selection using machine learning methods. In this approach, moderator selection is "automated" by implementing regression forest or similar algorithms to identify moderators substantially related to a summary effect estimate, while minimizing selection biases and multiple comparison concerns (Van Lissa, 2017).

with Y_i and ϵ_i defined as in (2) and X_{i1} ... X_{ip} indicating observed moderator variables included as covariates. The between-studies error component ξ_i is similarly defined as in (2), but here the variance of this quantity across all primary studies represents *residual* between-study variation, i.e., variation not explicitly modeled by the covariates (moderators) in the meta-regression. Note that the population average effect μ is replaced by the terms representing the covariate effect(s), with the model intercept term β_0 interpreted as the average program effect when all moderator values equal zero.[7]

Meta-regression has come into broad use for moderator analysis in aggregate data meta-analysis, principally because it offers several advantages over subgroup analysis. First, as noted above, it resolves the limitation of use of a single moderator variable. Second, its implementation is analogous to multiple linear regression in primary research in several ways: moderator variables can be entered as interactions and higher-order terms (e.g., cubic or quadratic), moderators can serve as "control" variables to adjust for sources of variance (here, between-study variation) that are not of substantive interest, and key model statistics (e.g., R^2) have relatively similar interpretations. As a result, meta-regression is seen as a fairly accessible technique to applied researchers, and of equal importance, to consumers of meta-analysis. Despite these strengths, meta-

[7] Meta-regression and subgroup analysis are conventionally distinguished by the use of continuous (or a mix of continuous and categorical) covariates in the former, and the use of a single categorical covariate in the latter. Analytically, however, both methods may be seen as extensions of the generalized linear model to meta-analysis, in a manner akin to multiple regression and analysis of variance (ANOVA) in primary data analysis (Nelson & Zaichkowsky, 1979; Thompson & Higgins, 2002). Additionally, and with some similarity to ANOVA and regression methods, subgroup analysis gained early popularity in meta-analysis, while meta-regression has only come into widespread use in the most recent two decades, likely because of the greater analytic flexibility of the approach (Tipton et al., 2019). As such, differences in nomenclature to some degree reflect the historical development of the methods, alongside a broadening of the use of meta-analysis to research contexts for which moderation by continuous (or multiple) covariates was of interest.

regression has several known limitations. As noted above, when many moderators are tested, multicollinearity can occur (Berlin & Antman, 1992; Hedges et al., 2010). Intuitively, it is possible that multicollinearity may be more likely when conceptually similar moderators, such as program implementation factors, are included together. A related concern is that multicollinearity can reduce statistical power to detect significant moderation effects in meta-regression, which compounds a lack of power arising from the small effective sample size (i.e., the number of studies) typical in meta-analysis compared with primary research (Baker et al., 2009).

Outside of the statistical considerations, another set of issues arises with regard to meta-regression model interpretation. The output of meta-regression with multiple moderators is informative about the magnitude of association between the moderators and the summary effect estimate, but as with multiple regression more generally, the joint distribution of moderators is not modeled, meaning that individual moderator coefficients must be interpreted while all other moderator effects are held constant. Although model-based predictions can be generated by specifying levels of the included moderators, such a process offers no insight into how probable each configuration of values is in practice (practice as represented by the various implementations of the program among the studies sampled). It can also be a laborious procedure when many moderators are of interest, and requires prior knowledge about reasonable levels of moderators to input. Thus, with meta-regression it becomes challenging to ascertain the substantive interrelation (joint distribution) of the moderators, i.e., how levels of the moderators co-occur in primary studies, and how that co-occurrence influences the magnitude or direction of the overall program effect.

The second interpretive consideration is that in meta-regression, moderators are generally treated as fixed effects (hence, the mixed-effects description of the model). This constraint implies that, while program effects themselves are allowed to vary across primary studies, the effect of a moderator on program effects is constant. If a program were implemented with different durations both across and within studies, for example, there may be interest in assessing whether longer programs had greater effect. To examine if such a disparity in effect exists, a meta-regression model with a continuous covariate indicating average program duration in minutes is fitted, finding that for each minute increase in average program duration, the outcome of interest is increased by one unit on average (here, unit increases are in the desirable/preventive direction of effect). This moderation effect is certainly informative, yet it is fixed across all studies regardless of whether the true individual study effects may still vary considerably. Individual study effects may range from, for instance, four to 14 units of effect. The expected increase in effect, as estimated by the meta-regression model, might lead to the conclusion that a unit of effect may be gained by both the lowest- and highest-effect studies, when in reality, it may be the case that in the highest-effect studies, an additional minute of the program makes no difference in effect (the program may already be comparatively long) while in the lowest-effect studies, increasing the average duration of the program may garner substantial increases in effect (perhaps because the program was too brief).

The above scenario is a form of aggregation bias sometimes referred to as *ecological bias*, and is visually represented in Figure 1. In this figure, the solid line represents the overall effect of average program duration across all studies from the meta-regression model. The slope of this line corresponds to the one-unit increase in effect for

each unit (minute) increase in intervention duration. Each dot represents a primary study (and its effect size), and the dashed lines indicate the moderation effect of program duration on program effect *within* each study. Note that such an analysis is not possible with aggregate data alone, but could be carried out with individual participant data if available. It can be observed that, while the meta-regression slope suggests that increased duration is associated with greater program effect, the relation of duration with effect at the study level is considerably more complex: in some studies the effect is minimal or absent, in others it is quite large and positive, and finally, in some studies the slope is negative, suggesting that longer program duration *reduced* the effect of the program.

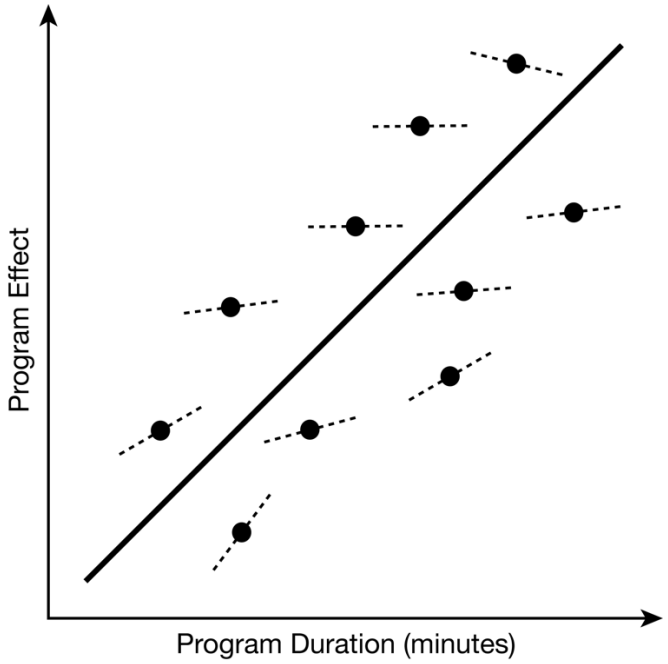

Figure 1. Representation of aggregation (ecological) bias in meta-analysis. Solid black line indicates output (slope) of meta-regression model, denoting a one-unit increase in program effect for each unit (minute) increase in program duration. Dots represent individual studies, and dashed lines indicate the relation (slope) of duration to effect within each study. Figure adapted from Baker et al. (2009) and Thompson (1994).

Aggregation bias can be particularly problematic when included moderators correspond to attributes of primary study samples, such as average participant age, or racial/ethnic or sex/gender composition (e.g., proportion female). In the context of meta-regression, aggregation bias often arises when the value associated with a moderator is a summary measure. In the cases of program duration or participant age, these values are likely to range across participants, so the point estimate needed for meta-regression may then be a mean of that range. This value could then be unrepresentative of the distribution of within-study values, for instance if participants in a study have a mean age of 30 years when individual ages range from 18–70 years. In this scenario, a participant who is 30 years of age is likely clinically and behaviorally different from a 70-year-old participant, and conclusions drawn from the use of a value of 30 years for the study in the meta-regression model may not accurately reflect the relation of age with program effect within that study. In contrast, use of a moderator whose value is uniform across all primary study participants (e.g., study-level attributes such as whether program delivery was monitored for fidelity) does not pose the same risk of aggregation bias. As suggested above, one solution to aggregation bias is to utilize individual participant data to more accurately model within-study moderation effects. Individual participant data is, however, resource-intensive to collect and can be challenging to analyze, and may be available in systematically different ways than aggregate data (and thus may represent a different population of studies and participants).

From an interpretation standpoint, the inferential risk accompanying aggregation bias may be particularly prominent when moderators are interpreted in relative isolation, for instance when concluding, in the above example, that program duration increases

program effect. A similar (mis)interpretation could be made with a participant attribute moderator such as age: because a program was found to be more effective among older participants across studies, it is assumed to be more effective among older participants within studies. It may be the case that with both of these moderators, in the absence of individual participant data, a more informative assessment of their relative influence on program effectiveness could be ascertained by considering the interrelation of multiple conceptually- or theoretically-related moderators concurrently. Inference, then, would be based on groups of studies sharing a set of (participant) characteristics, rather than driven by the linear relation of a single study's summary value of a moderator with program effectiveness. For example, if there were a known greater risk of an outcome of interest among older, female-identifying persons of non-white races/ethnicities, a multivariate analysis that concurrently utilized average participant age, proportion of female participants, and proportion of racial/ethnic minority participants, may be well suited to examine whether a program had a larger effect when implemented among multiple study samples with high likelihood of being older, non-white, and female-identifying compared to implementations among multiple study samples likely to be, on average, younger, white, and male-identifying. While such an analysis approach may not eliminate the risk of aggregation bias, it may offer insight into how studies differ in their population of focus, and as such provide information for future research about the populations to which a program could be generalized.

Primary Study Implementation Characteristics as Sources of Effect Size Variation

The influence of program implementation characteristics on program effectiveness is investigated in the research domain of Implementation Science (IS). In IS

frameworks such implementation factors are frequently represented as *drivers* of program effectiveness, in that they are conceptualized as being directly linked to how successfully a program is installed and delivered over time (Sims & Melcher, 2017). For example, one commonly identified set of drivers relates to the competency of program providers. Competency drivers include the selection of appropriate providers and the provision of adequate training to selected providers to ensure programs are delivered with fidelity to the original program protocol or design. Additional key implementation drivers include setting or contextual characteristics that facilitate effective program delivery, such as the availability of staffing, technological aids, and other support resources that increase the quality of program delivery by reducing provider workload or competing demands (Sims & Melcher, 2017).

Importantly, while some implementation characteristics elaborated by IS research, such as implementation support, fidelity monitoring, and provider training, have obvious relevance to programs implemented in controlled settings, the focus of IS research is chiefly on the successful dissemination and implementation of programs whose efficacy has already been determined in randomized trials or controlled evaluations. In this context, the emphasis is on how a program can be effectively delivered outside of a controlled research environment by community-based organizations or health systems, and by practitioners or peers rather than research staff. As a result, the applicability of IS frameworks to an earlier phase of research – namely, randomized trials to assess program efficacy – may be limited. For example, the presence of fidelity monitoring in a real-world implementation setting may relate to greater program effectiveness given the lower level of implementation control and the likelihood that providers may poorly deliver the

intervention when encountering fatigue, distractions, or implicit biases. In a controlled trial, however, fidelity monitoring may be one aspect of an overall more rigorously designed study that, by contrast, finds *lower* program effect (i.e., studies with greater fidelity monitoring may also have stricter inclusion/exclusion criteria, fewer potential biases in outcome assessment, and a more robust analytic approach). Despite this consideration, there is still value in examining the role of implementation factors in program efficacy in a controlled trial context, and one potential strategy to do so may involve examining an implementation factor alongside other design and methodological characteristics found across numerous implementations of a program. Such data – on multiple implementation characteristics of several or many instantiations of a program – can be readily available in a meta-analysis.

Motivating Example: Brief Substance Use Interventions

Brief substance use interventions (BIs) are low-resource interventions intended to reduce problem behaviors, such as alcohol or drug misuse, and in the healthcare setting are typically delivered in one session by a clinician, nurse, or behavioral health specialist (Parr et al., 2019). The structure and content of BIs are influenced by theoretical and conceptual models, such as motivational interviewing (Miller & Rollnick, 1991) and the transtheoretical model of behavior change (Prochaska & DiClemente, 1984), which emphasize the importance of personal agency in behavior change and suggest that interventions should strengthen individuals' self-efficacy and ability to alter harmful behavior patterns, rather that solely aim to deliver prescriptive or corrective guidance. This theoretical stance, which may also be termed "patient-centered" (Van Voorhees et al., 2009), is operationalized through a number of commonly-used BI components

including personalized normative feedback (information on levels of alcohol or substance consumption based on individually-appropriate limits and local or national statistics), goal-setting activities, decisional balance exercises (during which individuals self-identify positive and negative aspects of their use), booklets or information sheets, skills training (such as how to avoid drinking excessively in a social setting), and referrals to community support services (Tanner-Smith et al., 2020; Tanner-Smith & Lipsey, 2015). Some BI implementations have also included prescriptive advice as a primary or supplemental component, despite the apparent contradiction with a motivational approach. In these instances, advice-based BIs are argued to require less time and to better align with medical providers' traditional approach to patient care (authoritative advice to influence patient behavior), and consequently are viewed as more likely to be taken up by providers compared to motivational techniques (Davis et al., 2011; Miller & Rollnick, 2002; Van Voorhees et al., 2009).

Brief Intervention Characteristics and Components as Sources of Effect Size Variation

As noted above, BIs are typically single-session and delivered by a healthcare provider, but they may alternatively be composed of multiple sessions, and the content, provider profession, and delivery context of the BI may also vary. In particular, the number and types of components may differ among implementations of BIs as result of differences in theoretical rationale, trial or setting resources, desired level of control, or participant baseline severity level. In some cases, for instance, a BI may feature a single component, such as advice or normative feedback, while in other implementations, several components are included. Further, delivery of one or several components may be required by the intervention protocol, or alternatively, the provider is given some level of

discretion to choose from a set of components they deem most appropriate for a patient's severity, level of responsiveness to intervention, or readiness to change.

In addition to differences in components used, because BIs are generally intended to be responsive to patient needs and because provider demands may vary, BIs routinely differ in duration (both within a study and across studies). They may also be delivered in an in-person format or via a computer or tablet, by telephone, or using a pen-and-paper modality, and the delivery setting may be, for example, an emergency department, primary care office, hospital inpatient or outpatient environment, or university-based healthcare facility. Finally, the provider of the BI may be a physician, nurse practitioner/physician's assistant, nurse, behavioral health provider, social worker, or peer health worker, among other professions, and such providers may be clinical staff already working in the study setting or staff employed and installed in the setting by the research program.

Several primary studies and meta-analyses have found generally positive evidence of BI effectiveness among adults and adolescents (e.g., Kypri et al., 2008; Tanner-Smith & Lipsey, 2015; Vasilaki et al., 2006), yet it is conceivable that the magnitude of effect varies as a result of differences in primary study implementation characteristics and in the structure and delivery of the BI itself (Tanner-Smith & Lipsey, 2015). Indeed, prior meta-analyses have examined a number of implementation factors and BI characteristics using established moderator analysis approaches, including meta-regression. These analyses have found effect moderation by BI modality used (e.g., whether the BI used a motivational or information-only technique) and by specific BI components (e.g., whether goal-setting or decisional balance activities are administered, or information on

use consequences is provided) (Tanner-Smith & Lipsey, 2015; Vasilaki et al., 2006). Nevertheless, as described earlier, a given moderator of BI effectiveness may not be best considered in isolation, as such factors instead interrelate with other aspects of study implementation or intervention delivery. As such, there may be value in considering the co-occurrence of several potential moderators of BI effectiveness across numerous examples of BI implementation.

Analytic Data

Data analyzed here are drawn from an ongoing meta-analysis (k = 124) of randomized trials examining the effectiveness of brief interventions (BIs) for reducing alcohol and drug use behavior, as well as behavioral and health consequences of use, in general healthcare settings (Tanner-Smith et al., 2020). For the analyses presented in Chapter IV, synthesized effect sizes include those in the following domains:

1) drug use (cannabis, cocaine, methamphetamine, tobacco, other specific substance, or mixed drugs),
2) alcohol and/or drug use consequences ("use consequences") (arrests, driving-under-the-influence citations, or other criminal justice-related consequences; employment consequences, relationship consequences; sexual behavior consequences; health consequences; or other specific consequences).

Effects sizes were defined as a bias-adjusted standardized mean difference (Hedges' g) between groups receiving the BI and those in a control condition (e.g., treatment as usual, general health information, or sham intervention). Effect sizes reported as odds ratios were transformed to the standardized mean difference metric using a Cox transformation (Sánchez-Meca et al., 2003). When primary studies reported

multiple effect sizes within an outcome domain (e.g., cocaine and cannabis use within the drug use domain), effects sizes and their variance estimates were pooled using the approach described by Borenstein et al. (2009, p. 228). Meta-analysis models, described in detail in the Method section, were fitted separately for each outcome domain (drug use or use consequences). Moderators of primary interest are those representing study implementation characteristics, aspects of BI design and delivery, and participant sample attributes. Table 1 presents all moderators examined and related descriptive statistics. Some moderator levels were not represented in the data or were collapsed for analyses due to sparseness, and these instances are denoted in Table 1. As a secondary use of deidentified, aggregate data, the present analyses did not require institutional review board approval or oversight. The ongoing meta-analysis from which data were drawn was reviewed by the University of Oregon Institutional Review Board and deemed non-human subjects research.

Table 1. Moderator (indicator) variables used in analyses. Moderator levels not represented in the data are indicated by an asterisk (*) and levels collapsed due to sparseness are noted with a dagger (†).

Moderator	Levels	*k* (%)
	Efficacy-to-Effectiveness Staging	
Patients and problems	1. Clinical: patients presenting with typical/wide range of problems	14 (11.8)
	2. Mixed: routine patients paid for participation	105 (88.2)
	3. *Research: study-solicited volunteers	–
Practice context	1. Clinical: community setting with limited control	77 (65.3)
	2. †Mixed/Research: controlled research/university setting or mixed	41 (34.7)

Table 1. (continued).

Moderator	Levels	*k* (%)
Practitioners and therapists	1. Clinical: practicing providers	45 (38.1)
	2. Mixed: recruited clinicians (practicing providers paid for participation)	27 (22.9)
	3. Research: contracted non-clinicians/clinicians in training	46 (39.0)
Intervention context	1. Clinical: briefer/more realistic intervention duration/complexity	99 (83.9)
	2. Research: long and/or complex intervention	19 (16.1)
Therapeutic flexibility	1. Most flexibility: provider has full discretion over which intervention components to deliver	6 (5.1)
	2. Some flexibility: intervention is manualized but provider can tailor feedback based on patient severity/risk	96 (81.4)
	3. Little/no flexibility: strict adherence to protocol/script	16 (13.5)
Pre-therapy training	1. Clinical: brief training delivered in typical CE format	55 (46.6)
	2. †Mixed/Research: Full-day offsite for primary care staff, extensive/intensive training, required formal qualification	63 (53.4)
Intervention support	1. Clinical: implementation is supported with standard clinical resources	27 (22.9)
	2. Research: level of support not typically available in clinical setting (e.g., additional support staff during study, researcher assistance for intervention delivery/monitoring)	91 (77.1)
Intervention monitoring	1. Clinical: non-invasive monitoring (e.g., provider completed brief intervention summary after intervention visit)	52 (44.1)
	2. Research: invasive/intensive monitoring (e.g., direct observation, recording, ongoing/immediate feedback)	66 (55.9)
	Study Characteristics (Risks of Bias)	
Random sequence generation	1. Low	80 (64.5)
	2. †High/Unclear	44 (35.5)

Table 1. (continued).

Moderator	Levels	*k* (%)
Allocation concealment	1. Low	60 (48.4)
	2. †High/Unclear	64 (51.6)
Assessor blinding	1. Low	7 (5.6)
	2. †High/Unclear	117 (94.4)
Incomplete data	1. Low	41 (33.1)
	2. †High/Unclear	83 (66.9)
Selective reporting	1. Low	26 (21.0)
	2. †High/Unclear	98 (79.0)
Missing data handling [a]	1. MI/FIML	24 (20.0)
	2. LOCF/sensitivity analysis	31 (25.8)
	3. Listwise deletion or unclear	65 (54.2)
Reporting modality	1. Biological	10 (8.4)
	2. Interview	76 (63.9)
	3. Self-administered	33 (27.7)
Implementation monitoring	1. Not reported	52 (42.3)
	2. Reported	71 (57.7)
Implementation problem	1. †Reported or coder-identified	38 (30.9)
	2. Not reported	85 (69.1)
	Intervention Components [b]	
Prescriptive advice	1. No	45 (37.2)
	2. Yes	76 (62.8)
Information booklet	1. No	53 (43.8)
	2. Yes	68 (56.2)
Decisional balance exercise	1. No	84 (69.4)
	2. Yes	37 (30.6)
Goal-setting activity	1. No	61 (50.4)
	2. Yes	60 (49.6)
Normative feedback	1. No	33 (27.3)
	2. Yes	88 (72.7)
Skills training	1. No	104 (86.0)
	2. Yes	17 (14.0)
Referral	1. No	88 (72.7)
	2. Yes	33 (27.3)
	Intervention Characteristics	
Duration (min.)	Range: 2–124	–

Table 1. (continued).

Moderator	Levels	*k* (%)
	Sample Characteristics	
Proportion non-Hispanic white	Range: 0.00–0.99	–
Proportion female-identifying	Range: 0.00–1.00	–
Mean age (years)	Range: 14.91–69.16	–

[a] MI = multiple imputation; FIML = full-information maximum likelihood; LOCF = last observation carried forward. [b] Homework exercise, video doctor, and website access BI components were poorly represented in the data, and therefore were not included in analyses.

CHAPTER III

METHOD

In the prior section, several limitations of subgroup analysis and meta-regression for investigating moderators of program effectiveness in meta-analysis were outlined. They included, primarily, modeling challenges when numerous moderator variables are of interest, limitations relating to the interpretation of model output when multiple moderators are included, the inability to examine the co-occurrence of moderators in studies included in a meta-analysis, and the risk of aggregation or ecological bias depending on the nature of moderator variables (i.e., whether they represent fixed conditions across all primary study participants or summarize a range of values distributed among participants). To address some of these limitations, in the current section a finite mixture modeling-based methodology for moderator analysis is described. In this approach, primary studies are first grouped based on their probability of membership to subgroups characterized by multiple implementation, intervention, or sample attributes; following this classification procedure, random-effects meta-analysis can be carried out within each multivariate class of studies while accounting for imprecision in study classification. Thus, program effectiveness may be assessed, or secondary moderation analyses conducted, after studies are first characterized using a number of moderators of interest.

Overview of Finite Mixture Modeling

Finite mixture modeling encompasses a broad array of analytic approaches unified by the assumption that the overall population distribution of variables of interest are more accurately characterized by multiple – a mixture of – component distributions,

rather than a single distribution (McLachlan et al., 2019; McLachlan & Peel, 2000). A primary benefit of mixture modeling arises in cases when values exhibit, or are expected to exhibit, substantial heterogeneity (i.e., when those values may be distributed in qualitatively distinct ways and not simply differ in magnitude). More specifically, mixture models are optimal when modeling all values as a single distribution would obscure meaningful (i.e., informative) heterogeneity in those values (B. O. Muthén, 2001). In the lesser extreme, the impact of disregarding such heterogeneity or variation is that overall inference remains accurate but is incomplete: providing less insight than it could otherwise have, had underlying variation been more fully modeled. In the worst extreme, treating all values as drawn from a single distribution leads to fundamentally incorrect inference due to ignoring a substantial amount of variation. In such a case, an overall summary statistic of the single distribution masks underlying variation to the degree that the summary measure is unrepresentative of the underlying distribution(s).

A form of finite mixture modeling that has gained increased use in studies of social, behavioral, and public health outcomes is latent class analysis (LCA; Collins & Lanza, 2010; Goodman, 1974; Lazarsfeld & Henry, 1968; Masyn, 2013; McLachlan et al., 2019; McLachlan & Peel, 2000). The aim of LCA is to identify unobserved clusters or subgroups among observations (typically individuals). These subgroups are often referred to as classes, and are defined or characterized using multiple *indicator* variables thought to be informative about the latent structure of the data at hand. Indicator variables may be categorical or continuous,[8] and when modeled together, differentiate observations

[8] When continuous indicator variables are used, the analysis is sometimes referred to as latent profile modeling. Given the statistical and interpretive similarities, and for conciseness, modeling with categorical or continuous variables – or a combination – is here collectively referred to as latent class analysis (LCA).

into classes based on their observed pattern of values across indicators. A common application of LCA has been examining co-occurrence of substance use behaviors (see Collins & Lanza, 2010). In such analyses, response values to a set of indicators of use or non-use of several substances (e.g., alcohol, tobacco, cannabis, and cocaine) are modeled. The latent class model selected as best fitting then provides an estimate of the number of subgroups or classes that are likely to underlie the set of observations, and the probabilities of response to each level of the indicator variables within each class. Thus, for instance, one class may be characterized by a high probability of endorsing use of all substances; this class may be described as a "severe use" class, given that individuals belonging to the subgroup are likely to use all substances of interest. Conversely, the model may identify another class in which there is a low probability of use of each substance; respondents in this class might be described as "low risk". Importantly, while most applications of LCA implicitly define observations as persons such as in the above example, this is not a requirement of the underlying statistical model (Lazarsfeld & Henry, 1968, p. 17).

Model Estimation and Class Enumeration

Two sets of parameters are of interest in LCA, and are typically estimated using a maximum-likelihood approach and an expectation-maximization algorithm (Dempster et al., 1977; see also Collins & Lanza, 2010; B. O. Muthén, 2001). The first set of parameters is the conditional item response probabilities, which estimate the probability of endorsing each value of an indicator within each class, and the second set is class probabilities, which give the proportion of observations in each class. If a vector containing a full pattern of values across indicator variables is defined as $\boldsymbol{Y}$, and a

specific response pattern as $\boldsymbol{y}$, then the probability of that response pattern is estimated by

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{t} P(C = t) P(\boldsymbol{Y} = \boldsymbol{y} | C = t)\,, \tag{8}$$

with C indicating the latent class variable composed of multiple latent classes t ($t = 1, 2, \ldots, T$).[9] Here, $P(C = t)$ is the probability of an observation belonging to class t, and $P(\boldsymbol{Y} = \boldsymbol{y} | C = t)$ is the probability of a particular response pattern occurring in class t. Taken together, the two sets of parameter estimates provide insight into the latent categorical structure of the available data when the model with the correct number of classes has been selected and classes are well-separated (i.e., item response probabilities indicate distinct response or value patterns in each class). In the absence of accurate enumeration and sufficient class separation, characteristics of classes including the proportion of observations belonging to each class and the probabilities of indicator values in each class are of little substantive use. Finally, the values on indicator variables are typically assumed independent conditional on class membership.

A number of approaches to selecting the best fitting (i.e., accurately enumerated) latent class model have been examined (for review, see Nylund et al., 2007). First, information criteria such as the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) are frequently used for model comparison and selection. The BIC metric has been found to be well-performing across a

[9] In LCA literature, individual classes are sometimes denoted as k. This usage is not implemented here, and this term is instead represented by t to avoid confusion with the definition of k as the number of primary studies in a meta-analysis. This usage is virtually ubiquitous in meta-analytic literature, and is maintained here.

variety of mixture applications and modeling scenarios, and is generally recommended (Cheung, 2008; B. O. Muthén, 2001). When comparing two models using BIC, the model with the smaller value of the criterion is considered better fitting. Alternatively, likelihood ratio tests (LRT) can be employed to assess the statistical significance of differences in model likelihood. In particular, an adjusted version of an LRT (aLRT) for models with differing class structures was proposed by Lo, Mendel, and Rubin (2001), and has been found to be well performing for latent class model selection (Lubke & Muthén, 2005). When using an aLRT, a model with t classes is compared to a model with $t + 1$ classes; a significant p-value indicates that the larger model is better fitting than the smaller model, while a non-significant p-value suggests that the larger model fits no better than the smaller model (and thus the smaller model is more parsimonious). Importantly, model selection is best carried out using multiple statistics, such as a combination of BIC and an aLRT, to corroborate model fit (Cheung, 2008; Collins & Lanza, 2010; Masyn, 2013).

Examining Class Membership

A critical aspect of LCA is quantifying class membership, which involves identifying those observations likely to belong to each class based on their indicator variable values. Accurately characterizing class membership is particularly important when there is an interest in auxiliary analyses: testing whether other variables (covariates) predict membership in classes, examining whether class membership predicts later outcomes, or conducting secondary inference within classes (i.e., fitting additional models among observations within each class). Estimating class membership is carried out using Bayes' theorem,

$$P(C = t|\boldsymbol{Y} = \boldsymbol{y}) = \frac{P(C = t)P(\boldsymbol{Y} = \boldsymbol{y}|C = t)}{P(\boldsymbol{Y} = \boldsymbol{y})}, \quad (9)$$

which provides the posterior probability of each observation's membership to each class given the observation's pattern of values across the indicator variables ($\boldsymbol{y}$). Classes are mutually exclusive and exhaustive, such that an observation's posterior probabilities of membership to classes sum to one (Collins & Lanza, 2010).

While interpretation of posterior probabilities is fairly straightforward, it must be recognized that classes are unobserved, and therefore assignment to classes is both probabilistic and, to a measurable extent, uncertain.[10] To understand the source of this uncertainty, it is helpful to observe that Equations 8 and 9 pertain to observations' patterns of values on indicator variables, not to individual observations themselves (which would be denoted with an i subscript). Thus, while individual observations may only belong to one class, and in a well-defined latent class model, a given response pattern uniquely characterizes each class, observations with the same response pattern may belong to *different* classes because of the unobserved nature of the classes (i.e., because there is error associated with measurement and classification). As a result of this classification uncertainty, when using class membership probabilities in auxiliary analyses it becomes important to incorporate the uncertainty of membership assignment into those analyses in order to provide accurate estimates of standard errors and statistical significance.

[10] The quality or precision associated with classification is termed *entropy*, and can range in value from 0 (classification no better than random chance) to 1 (perfect classification) (Masyn, 2013).

Various methods have been proposed for carrying out auxiliary analyses in latent class models while accounting for classification uncertainty (for further discussion, see Asparouhov & Muthén, 2014; Bakk et al., 2014, 2016; Lanza et al., 2013; Vermunt, 2010). In general, the methods vary in how each addresses two central issues. The first, and most directly related to class membership estimation, is the concern mentioned above regarding carrying class assignment uncertainty into secondary analyses. The second issue is ensuring the estimation of class structure is not influenced by variables to be used in auxiliary analyses. When the latter occurs, covariates intended to predict, or be predicted by, latent classes instead become indicator variables, attenuating estimates of the relation between the covariates and the latent class variable. This problem is not directly related to uncertainty in how class membership is derived, but has influenced how the issue is addressed. Namely, a multistep procedure has been developed in which 1) the latent class structure is first characterized without auxiliary variables in the model (avoiding the second issue), 2) class membership is assigned and uncertainty in the assignment estimated, and 3) secondary analyses that incorporate the uncertainty quantified in the second step are carried out (addressing the first issue). Several of the methods for auxiliary analysis in LCA make use of such a multistep approach, though with varying degrees of success in ensuring the estimation of class structure is not influenced by auxiliary variables and that standard errors are correctly estimated (see Bakk et al., 2014).

One approach that has been found to generally achieve both aims and that is versatile with regard to secondary inference is the so-called BCH method, first proposed by Bolck, Croon, and Hagenaars (2004) and subsequently refined by Vermunt (2010). In

the BCH method, the first step as described above is conducted, generating item response probabilities and class proportions using Equation 8 and class membership probabilities via Equation 9. The class membership (posterior) probabilities provide the individual observations' predicted class membership W, with specific predicted classes defined as s ($s = 1, 2, \ldots, S$). The more precisely the model predicts class membership (i.e., the closer the posterior probabilities are to one and zero), the more certainty there can be that $s = t$, with t as defined above (Bakk et al., 2013). In the second step, the average classification uncertainty across all patterns of indicator values can be estimated as the probability that an assigned class (s) is a true class (t), or

$$P(W = s|C = t) = \frac{\frac{1}{n}\sum_i P(C = t|\boldsymbol{Y} = \boldsymbol{y_i})P(W = s|\boldsymbol{Y} = \boldsymbol{y_i})}{P(C = t)} \ . \qquad (10)$$

Bolck, Croon, and Hagenaars (2004) and Vermunt (2010) showed that the estimated classification error can be non-linearly transformed to generate observation-level weights that reflect both class assignment and its uncertainty. These weights can be incorporated into the third step, above, in effect creating a multiple-group analysis in which secondary relations can be examined, or auxiliary inference carried out, while incorporating classification error (Asparouhov & Muthén, 2018).

Prior Applications of Mixture Modeling in Meta-Analysis

To date, mixture modeling in the meta-analytic context has focused on fitting mixture models directly to primary study-level effect sizes as a strategy to address (but not predict or explain) between-study variation (e.g., Böhning, 2005; Schlattmann, 2009; van Houwelingen et al., 2002; Xia et al., 2005). The aim in these applications has been to cluster studies into groups within which effects are homogeneous. This strategy can be

useful when there is considerable between-study variation in effects among studies, but there is reason to believe this variation is random and unexplainable. It may also be helpful when there is explainable between-study variation in effects, but few or no moderator variables are available to examine as potential sources of that variation. A final existing application of mixture models in meta-analysis is in the context of synthesizing studies of diagnostic test accuracy. In these studies, effects group into a known number of distributions (bivariate, i.e., one for test sensitivity and one for test specificity), and mixture models are used to group studies into classes of differing accuracy (Eusebi et al., 2014; Schlattmann et al., 2015). In contrast to these approaches, the proposed method applies mixture modeling to potential effect moderators, rather than effect sizes themselves, with the aim of investigating the co-occurrence of moderators. The relation of class membership with program effects is then estimated to assess whether moderator co-occurrence meaningfully influences program effectiveness. The method, unlike prior applications, also does not treat study membership to classes with certainty, and instead incorporates classification uncertainty or imprecision into model inference.

Application

In the previous section, mixture modeling was overviewed as a versatile framework for multivariate inference. Latent class analysis, in particular, was described as an analytic approach useful for investigations of the relation among multiple variables that may define discrete classes in a set of observations. In the current section, the proposed methodology, which applies LCA to moderator analysis in meta-analysis, is described.

When a population of values can be (or is assumed to be) characterized by a single distribution, it can be conceptualized that those values belong to a single class. That is, if a mixture model were fitted to those values, a single class would sufficiently characterize their distribution. Thus, in the context of meta-analysis, a random-effects model (Equation 2) can be re-expressed with this implicit (latent) class variable C made explicit:

$$Y_i|C = \mu_c + \xi_{ic} + \epsilon_{ic} \ . \tag{11}$$

In this instance, C receives no subscript because there is, as described, only one class. It is clear, however, that C can be defined as $c_1, c_2, \ldots, c_T$; that is, as a categorical latent variable representing multiple latent classes t ($t = 1, 2, \ldots, T$). Classes would then reflect qualitatively distinct distributions of values (effects); the question then arises of how such classes might be characterized. As described in the prior section, latent subgroups or classes are defined by multiple indicator variables that provide information on the underlying structure of the available observations. The nature of these indicator variables (i.e., what they measure) characterizes the latent space to be modeled. Importantly, and in contrast to traditional subgroup analysis in meta-analysis, this latent space is multivariate. Indeed, the principal difference between Equation 6, which presents the random-effects model for subgroup analysis, and Equation 11, is that the traditional subgrouping variable G defines subgroups that are observed and univariate in nature, while the class variable C reflects subgroups that are latent and multivariate (i.e., derived from the joint distribution of indicator variables).

In the present application, primary interest is in utilizing indicators related to study implementation methodology, intervention design, and sample characteristics to define the categorical latent space. Within these categories (classes), primary studies are classified using the BCH method and summary statistics estimated, including within-class summary effect size estimates $\bar{Y}|C$,

$$\bar{Y}|C = \frac{\sum_{ic=1}^{kc} W_{ic} Y_{ic}}{\sum_{ic=1}^{kc} W_{ic}} \tag{12}$$

and estimates of the within-class magnitude of between-study variation (i.e., τ_c^2). Importantly, if the co-occurrence of indicator variables does moderate effects, τ_c^2 provides a measure of *residual* between-study variation, or the magnitude of variation remaining among effect sizes in each class after some part of the total between-study variation is accounted for by the moderators defining the classes.

Initial Stage: Estimating Multivariate Classes of Primary Study Characteristics

The approach just described was implemented using the motivating data by first estimating latent class models with the selected indicator (moderator) variable values for all primary studies. At this stage, comparative latent class model fit was evaluated using the model BIC value (Schwarz, 1978) and the adjusted likelihood ratio test (Lo et al., 2001), described above, as well as assessment of model entropy, class separation, and overall model interpretability. Final models were fitted with 400 random starts followed by a second fitting with 800 random starts to ensure model likelihood was replicated and to reduce the risk that the identified model was at a local maximum. When supported by the data, alternative models were estimated (e.g., with one additional class than was suggested by the fit statistics). The rationale for this approach was that, as exploratory

analyses, differing class structures could provide additional information about the interrelationship of moderators and their impact on the overall estimate of BI effectiveness within classes. Further, because in the meta-analytic context the number of observations (studies) is small compared to most modeling settings in which participants constitute observations, the results of the likelihood ratio test are subject to small-sample biases and power concerns, and therefore may not provide a definitive indication of the best-fitting model (Masyn, 2013). Finally, whether an individual indicator was highly discriminating between classes (i.e., was associated with response probabilities approaching zero or one) was not used as a criterion for inclusion or exclusion of the variable from a model. Rather, when an indicator was found to be poorly discriminating (in an otherwise well-fitting model), the variable was maintained in the final model as it provided some information about the moderator's role in differentiating among program effects (or the absence of such a role).

Moderator variables (Table 1) were included in specific latent class models based on their topical relatedness (e.g., pertaining to characteristics of the BI, such as intervention components) or because they represented elements of an existing scale or assessment methodology. Examples of the latter include the Cochrane Risk of Bias Tool (Higgins et al., 2011), which assesses aspects of a study design or implementation such as quality of randomization and degree of allocation concealment that may bias reported effects, and measures of the research stage of various elements of the original trial (i.e., whether characteristics of the study were representative of efficacy or effectiveness testing; Kaner et al., 2003). Moderator values were populated during the original data extraction of the motivating systematic review and meta-analysis by two independent

coders, whose inputs were reconciled as needed by a third coder. Missing moderator values were not imputed to avoid obscuring any latent groupings in the data.

Second Stage: Conducting Random-Effects Meta-Analysis Within Classes

Once a model was selected, random-effects meta-analysis models were estimated within each class, as defined by BCH weights calculated during the fitting of the selected latent class model. Synthesized effect sizes are those available from studies assigned to each class. Because mixture modeling, including LCA, is a latent variable method, meta-analyses carried out using the proposed approach was conducted in an SEM framework. Cheung (2008, 2013, 2015) showed that traditional fixed-, random-, and mixed-effects meta-analytic models could be estimated using SEM, and at the same time, benefit from modeling tools available in SEM.[11] These include, for instance, integrated missing data handling and robust statistics (Cheung, 2015). In the current application, a further rationale is that mixture modeling can be construed as a special case of SEM (Cheung, 2008), and as such, the LCA approach discussed here could be straightforwardly integrated with meta-analysis when the latter is also viewed as a specialization of SEM (Cheung, 2015). When implementing random-effects meta-analysis in an SEM framework, the focus of modeling is a latent random variable representing the true, study-level effect size, whose mean across studies is the sample estimate of the overall (average) population effect size (μ in Equation 2), and whose variance equals the magnitude of between-study variation in effects, τ^2. Central to the approach is

[11] Note that meta-analysis carried out using SEM is distinct from meta-analytic structural equation modeling (MASEM), which describes the synthesis of correlation matrices from primary studies (see Jak, 2015, for further discussion).

transforming sampling errors so they are identically distributed; doing so permits primary studies to be treated, in effect, as individual observations (Cheung, 2008). This transformation is accomplished by multiplying the terms of the random-effects model (Equation 2) by the inverse of the square-root of the variance, so that the error variance of each study-level effect size is transformed to a value of one.

For models presented in Chapter IV, random-effects models with the above transformation were estimated within each class as random-slopes models using MLE with robust standard errors (Bakk & Vermunt, 2015). Because estimates of τ^2 can be downwardly biased with the use of MLE (Viechtbauer, 2005), classes were monitored for instances when there were few effect sizes available for within-class meta-analyses, and limitations in interpretation arising from this scenario are identified and discussed as necessary. Model results for overall effect sizes and estimates of τ^2 are accompanied by 95% confidence intervals. All models were estimated using M*plus* version 8.4 (L. K. Muthén & Muthén, 2019), while data handling was carried out in R version 3.6.3 (R Core Team, 2020; RStudio Team, 2019).

CHAPTER IV

RESULTS

In the present section, results of fitted models are presented in the following fashion. First, the findings of the mixture (LCA) model for a particular set of moderators (e.g., those related to study efficacy-to-effectiveness staging) are described. Discussion of class structure is followed by interpretation of the meta-analytic findings, both for the overall (one-class) model and within-class models, for each of the outcome domains of interest (drug use and use consequences). For convenience of exposition and so that class structure and summary effect size estimates can be considered alongside one another, mixture model findings are re-presented for each outcome domain. In all tables in this chapter, boldfaced item response probabilities for mixture model findings are those greater than 0.60, and boldfacing of confidence intervals indicates statistical significance at $\alpha = 0.05$. Additionally, the number of studies within each class (k_c) is provided. Estimates of k_c are derived from posterior probabilities of class memberships, and as noted in the previous section, the BCH method incorporates classification uncertainty, and therefore class counts and proportions are approximate. The number of effect sizes available to synthesize in each class (k_{es}) is also estimated from posterior probabilities, and may not correspond to the number of studies in each class given that not all primary studies reported findings for all outcome domains. The number of effect sizes for each outcome domain is indicated in all results tables presented in this chapter, and in the meta-analytic results sections of the tables, an asterisk denotes instances when a parameter's standard error (and confidence interval) could not be estimated. Such an issue arises when a class has few effect sizes and the within-class estimate of between-

study variation in effects (τ_c^2) approaches zero, and this and other implications of class imbalance in available effect sizes are considered at length in Chapter V. Finally, unless otherwise noted, for the following mixture models a two-class solution was found to be best-fitting based on model fit statistics and consideration of parsimony and interpretability. As a result, model fit is not discussed in the present chapter, and detailed fit information – including BIC values, entropy levels, and likelihood ratio test p-values for all mixture models – is presented in the Appendix.

Efficacy-to-Effectiveness Staging

Rating study characteristics along a continuum from efficacy to effectiveness is a means of assessing the degree to which a study is representative of a highly controlled, research-typical study, or a minimally controlled, pragmatic or clinically-typical study design. In other terms, it can describe whether a study would be considered a test of a program's efficacy, in which case greater control is warranted to reduce spurious associations, or a test of its effectiveness, when there is greater interest in determining program impact in a realistic implementation setting. As such, when efficacy-to-effectiveness indicators are considered together, they provide some sense of the feasibility with which a program could be implemented in a real-world setting where there are typically fewer resources in the form of, for instance, staff to conduct rigorous intervention monitoring or implementation support.

Table 2 presents results of a two-class mixture model of efficacy-to-effectiveness indicator variables, with data provided by 119 primary studies. Studies in both classes were likely to have mixed or research-typical patients and problems (i.e., routine patients paid for participation or volunteers solicited for the study), and to have clinically-typical

Table 2. Two-class model of efficacy to effectiveness moderation of drug use effect sizes.

$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	
	Class 1	Class 2
Patients and Problems		
1 (clinical)	0.08	0.21
2 (mixed/research)	**0.92**	**0.79**
Practice Context		
1 (clinical)	0.56	**0.84**
2 (mixed/research)	0.44	0.17
Practitioners		
1 (clinical)	0.22	**0.71**
2 (recruited clinician)	0.31	0.07
3 (mixed/research)	0.47	0.22
Intervention Context		
1 (clinical)	**0.80**	**0.91**
2 (mixed/research)	0.20	0.09
Therapeutic Flexibility		
1 (low)	0.04	0.07
2 (moderate)	**0.88**	**0.68**
3 (high)	0.08	0.25
Pre-therapy Training		
1 (clinical)	0.25	**0.91**
2 (mixed/research)	**0.75**	0.09
Intervention Support		
1 (clinical)	0.03	**0.64**
2 (research)	**0.97**	0.36
Intervention Monitoring		
1 (clinical)	0.17	**1.00**
2 (research)	**0.83**	0.00
k_c	80	39
(%)	(67.4)	(32.6)
k_{es}	26	5
$\bar{g}_c$ [95% CI]	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ_c^2 [95% CI]	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]

interventions (i.e., lower complexity and/or briefer duration) delivered with moderate flexibility (i.e., manualized or protocol-driven intervention with some provider discretion on the type and number of components delivered). The first class (Class 1), composed of approximately 80 studies, was characterized by studies with a high probability of mixed or research-typical pre-therapy training, which may include especially long or intensive training for providers, or off-site training for primary care providers; a high probability of research-typical intervention support, which may include additional staff for risk assessment, intervention delivery, or administrative tasks; and a high probability of research-typical intervention monitoring, which may feature direct observation, immediate corrective feedback, or other invasive or resource-intensive monitoring approach. The second class (Class 2) was composed of the remaining 39 studies, and in contrast to Class 1, included studies that were likely to feature clinically-typical levels of pre-therapy training (i.e., brief training in the format of continuing education), intervention support (i.e., few to no additional support staff), and monitoring (i.e., limited or indirect monitoring, such as completing a brief report after BI delivery). In addition, studies in this class had a high probability of implementation in clinically-typical environments, such as a community-based setting, and of BIs being delivered by a practicing doctor, nurse, or other working provider rather than a research-recruited clinician, non-clinician, or trainee such as a graduate student.

The random-effects meta-analyses presented in Table 2 consider the effect of BIs on drug use, an outcome reported by 31 of the included studies. The overall (single-class) model indicates that the intervention significantly reduced drug use, $\bar{g} = 0.090$, 95% CI [0.038, 0.143], but with significant between-study variation in effect sizes, $\tau^2 = 0.010$,

95% CI [0.001, 0.019]. In the first class, which includes approximately 26 of the 31 available effect sizes and is generally characterized by research-typical implementation features, the BI effect is similar in both direction and magnitude to the overall model, $\bar{g}_c$ = 0.085, 95% CI [0.033, 0.137]. In this class, there remains the same amount of between-study variation in effect sizes, τ_c^2 = 0.010, 95% CI [0.002, 0.018], though the quantity is somewhat more precisely estimated. By contrast, in the class that reflected a more clinically-typical implementation, synthesis of the five available effects sizes suggested a much larger BI effect, $\bar{g}_c$ = 0.206, 95% CI [-0.078, 0.491], with increased between-study variation compared to the overall model and to the first class, τ_c^2 = 0.029, 95% CI [-0.080, 0.138]. Both estimates are nonsignificant, a finding likely the result of the limited number of effect sizes available in the class (an issue further discussed in Chapter V).

A similar pattern of results was found for the effect of BIs on use consequences (Table 3). Here, the BI had a larger overall effect compared to drug use, $\bar{g}$ = 0.106, 95% CI [0.063, 0.150], with a comparable estimate of between-study variation in effects τ^2 = 0.014, 95% CI [0.005, 0.022]. In the research-typical class (Class 1), the BI effect is again similar to the overall estimate, $\bar{g}_c$ = 0.099, 95% CI [0.054, 0.143], with equivalent between-study variation, τ_c^2 = 0.014, 95% CI [0.005, 0.022]. In the clinically-typical class, the BI effect was substantially larger than in the research-typical class, $\bar{g}_c$ = 0.310, 95% CI [0.157, 0.464], with no remaining between-study variation in effect sizes, τ_c^2 = 0.000, 95% CI [*]. Note that for the use consequences model, the Class 2 summary effect was significantly different from zero, while the standard error for τ_c^2 could not be estimated in this model.

Table 3. Two-class model of efficacy to effectiveness moderation of use consequences effect sizes.

$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	
	Class 1	Class 2
Patients and Problems		
1 (clinical)	0.08	0.21
2 (mixed/research)	**0.92**	**0.79**
Practice Context		
1 (clinical)	0.56	**0.84**
2 (mixed/research)	0.44	0.17
Practitioners		
1 (clinical)	0.22	**0.71**
2 (recruited clinician)	0.31	0.07
3 (mixed/research)	0.47	0.22
Intervention Context		
1 (clinical)	**0.80**	**0.91**
2 (mixed/research)	0.20	0.09
Therapeutic Flexibility		
1 (low)	0.04	0.07
2 (moderate)	**0.88**	**0.68**
3 (high)	0.08	0.25
Pre-therapy Training		
1 (clinical)	0.25	**0.91**
2 (mixed/research)	**0.75**	0.09
Intervention Support		
1 (clinical)	0.03	**0.64**
2 (mixed/research)	**0.97**	0.36
Intervention Monitoring		
1 (clinical)	0.17	**1.00**
2 (mixed/research)	**0.83**	0.00
k_c	80	39
(%)	(67.4)	(32.6)
k_{es}	41	7
$\bar{g}_c$ [95% CI]	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ_c^2 [95% CI]	0.014 **[0.005, 0.022]**	0.000 *

Study Characteristics (Risks of Bias)

As noted above, several characteristics of a primary study's design or implementation that may pose a risk of biasing estimates of program effect have been codified in the Cochrane Risk of Bias Tool (Higgins et al., 2011). Other potential factors that may influence reported effect sizes include missing data handling (e.g., use of a method which inflates bias, such as listwise deletion), whether self-report assessment or objective measures (e.g., assays for blood alcohol level) were used, and the presence or absence of fidelity monitoring (assuming a known link between adherence to an intervention's design or protocol and the intervention's effectiveness). Additionally, studies may also directly report (or coders may identify) implementation problems, such as difficulty recruiting participants or substantial differential attrition between groups.

Table 4 presents a two-class mixture model of primary study characteristic indicator variables, with data provided by 124 studies.[12] Studies in both classes were likely to have unclear or high risk associated with assessor blinding (e.g., outcome assessors unblinded to group assignment), and to have unclear to high risk of selective reporting (e.g., no published study protocol, reporting of outcomes that were not pre-specified, or failure to report pre-specified outcomes). All studies were also likely to use an interviewer assessment format and to report no implementation problems. Approximately 76 studies comprised the first class (Class 1); in this class, studies were

[12] Fit information for mixture models using study characteristic indicators suggested a two-class model was not better fitting than a single-class model (see Appendix). The two-class model is presented as exploratory. Additionally, a sensitivity analysis was conducted with only measures derived from the Cochrane Risk of Bias Tool (i.e., random sequence generation, allocation concealment, assessor blinding, incomplete data, and selective reporting). Findings of both the two-class mixture model and within-class meta-analyses did not substantively differ from those presented in Tables 4 and 5.

Table 4. Two-class model of study characteristic moderation of drug use effect sizes.

$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	
	Class 1	Class 2
Random Sequence Generation		
1 (low)	**0.82**	0.38
2 (high/unclear)	0.18	**0.62**
Allocation Concealment		
1 (low)	**0.63**	0.26
2 (high/unclear)	0.37	**0.74**
Assessor Blinding		
1 (low)	0.07	0.03
2 (high/unclear)	**0.93**	**0.97**
Incomplete Data		
1 (low)	0.47	0.11
2 (high/unclear)	0.53	**0.89**
Selective Reporting		
1 (low)	0.29	0.08
2 (high/unclear)	**0.71**	**0.92**
Missing Data Handling		
1 (MI/FIML)	0.32	0.00
2 (LOCF/sensitivity)	0.33	0.14
3 (listwise or unclear)	0.35	**0.86**
Reporting		
1 (biological)	0.06	0.12
2 (interview)	**0.63**	**0.66**
3 (self-administered)	0.31	0.22
Monitoring		
1 (no)	0.24	**0.71**
2 (yes)	**0.76**	0.29
Implementation Problem		
1 (yes/possible)	0.32	0.29
2 (no)	**0.68**	**0.71**
k_c	76	48
(%)	(61.2)	(38.8)
k_{es}	26	5
$\bar{g}_c$ [95% CI]	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ_c^2 [95% CI]	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]

likely to have low risk associated with random sequence generation (e.g., use of a computerized random-sequence generator) and allocation concealment (e.g., allocations conducted off-site and secured in sealed opaque envelopes). Additionally, these studies had a high probability of conducting fidelity monitoring. In the second class, composed of approximately 48 studies, studies were likely to be characterized by unclear or high random sequence generation risk (i.e., unclear randomization strategy, or use of any non-random allocation procedure), unclear or high allocation concealment risk (i.e., no reported concealment method, or use of any concealment approach that allowed participants, investigators, or other staff to foresee assignment), and unclear or high incomplete data risk (e.g., high attrition and/or use of an as-treated rather than intention-to-treat analysis approach). These studies were also likely to use a bias-inducing missing data handling strategy, such as listwise deletion, and to lack fidelity monitoring.

Findings of the overall meta-analysis of BI effect on drug use (the first section of Table 4), which suggest that BIs lead to a reduction in drug use, are identical to those presented in Table 2 and discussed in the prior section. Synthesis of the available effect sizes in each class also yielded similar findings. In the first class, the class-specific overall effect size, $\bar{g}_c$ = 0.085, 95% CI [0.033, 0.137], and between-study variation estimate, τ_c^2 = 0.010, 95% CI [0.002, 0.018], were again similar to the single-class estimate. In the second class, a much larger BI effect was found, $\bar{g}_c$ = 0.206, 95% CI [-0.078, 0.491], again with increased between-study variation compared to the overall model and to the first class, τ_c^2 = 0.029, 95% CI [-0.080, 0.138]. Findings for the use consequences outcome domain (Table 5) also paralleled those presented in Table 3.

Table 5. Two-class model of study characteristic moderation of use consequences effect sizes.

$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	
	Class 1	Class 2
Random Sequence Generation		
1 (low)	**0.82**	0.38
2 (high/unclear)	0.18	**0.62**
Allocation Concealment		
1 (low)	**0.63**	0.26
2 (high/unclear)	0.37	**0.74**
Assessor Blinding		
1 (low)	0.07	0.03
2 (high/unclear)	**0.93**	**0.97**
Incomplete Data		
1 (low)	0.47	0.11
2 (high/unclear)	0.53	**0.89**
Selective Reporting		
1 (low)	0.29	0.08
2 (high/unclear)	**0.71**	**0.92**
Missing Data Handling		
1 (MI/FIML)	0.32	0.00
2 (LOCF/sensitivity)	0.33	0.14
3 (listwise or unclear)	0.35	**0.86**
Reporting		
1 (biological)	0.06	0.12
2 (interview)	**0.63**	**0.66**
3 (self-administered)	0.31	0.22
Monitoring		
1 (no)	0.24	**0.71**
2 (yes)	**0.76**	0.29
Implementation Problem		
1 (yes/possible)	0.32	0.29
2 (no)	**0.68**	**0.71**
k_c	76	48
(%)	(61.2)	(38.8)
k_{es}	41	7
$\bar{g}_c$ [95% CI]	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ_c^2 [95% CI]	0.014 **[0.005, 0.022]**	0.000 *

The similarity in findings between efficacy-to-effectiveness and study characteristic models reflects the fact that the same effect sizes were categorized into the two classes identified by each model. This occurs because every study did not have an available effect size for the outcomes analyzed, so differences in the studies assigned to each class did not necessarily correspond to substantial differences in the effect sizes assigned to each class (i.e., some studies assigned to different classes between models did not have a corresponding effect size). Nevertheless, interpretation of the study characteristic models is distinct from that of the efficacy-to-effectiveness models. In the present models, the class having the smaller estimate of BI effect is characterized by studies having fewer risks of bias as well as fidelity monitoring, while the class having the larger estimate of BI effect is composed of studies with a higher probability of several risks of bias, including risk related to random sequence generation, allocation concealment, assessor blinding, and the absence of fidelity monitoring.

Intervention Duration

Beyond attributes of the primary study design and implementation, aspects of the BI itself may have some relation with its effectiveness. As noted in Chapter II, BIs can differ in their duration, as well as in the types and number of intervention components delivered to participants. Duration, typically reported in mean or median number of minutes, can represent the length of a single session, or less commonly, the total length of a multi-session BI. As a continuous quantity, duration may be examined using meta-regression to determine the influence of minute-increases in BI length on the overall effect size estimate. This said, there may be interest in whether primary studies, considered together, reflect groupings of effect size durations whose average values may

be more useful to decision-making regarding BI implementation than changes in effect by minute-units.

Table 6 presents the output of a two-class mixture model of BI duration, which utilized 103 primary studies providing BI duration information. Approximately 97 studies were assigned to the first class (Class 1), and on average, the duration of these studies was 21.2 minutes. In contrast, the average duration of the approximately six studies in the second class (Class 2) was considerably longer at 108.6 minutes. The meta-analytic findings indicate the estimated BI effect on drug use among studies in the first class, $\bar{g}_c$ = 0.091, 95% CI [0.035, 0.146], was similar in magnitude and significance to the overall (single-class) effect estimate, $\bar{g}_c$ = 0.090, 95% CI [0.038, 0.143]. The BI effect estimate in the second class, which was characterized by longer-duration interventions, was substantially larger, $\bar{g}_c$ = 0.204, 95% [-0.075, 0.482], but not significantly different from zero. In the first class, between-study variation was similar in magnitude to the overall model and remained significant, τ_c^2 = 0.011, 95% CI [0.002, 0.019], while in the longer-duration class, there was increased between-study variation, τ_c^2 = 0.028, 95% CI [-0.070, 0.126]. Results for the use consequences outcome domain (Table 7) bear some similarities to the drug use outcome domain, particularly in the first class where the BI effect and between-study variation estimates closely parallel the overall model estimates. In the second class of longer-duration BIs, the BI effect estimate is again substantially larger than the shorter-duration BI class; in this instance, however, the effect estimate remains significant, $\bar{g}_c$ = 0.323, 95% [0.130, 0.516], and the estimate of residual between-study variation is reduced to zero, τ_c^2 = 0.000, 95% CI [-0.145, 0.146]. Taken together,

these findings suggest that the use consequences effect sizes in the longer-duration class are a more homogeneous set of effects than the drug use effect sizes included in the longer-duration class presented in Table 6.

Table 6. Two-class model of brief intervention duration moderation of drug use effect sizes.

$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	
	Class 1	Class 2
Duration *Minutes (mean)*	21.2	108.6
k_c	97	6
(%)	(94.0)	(6.00)
k_{es}	24	5
$\bar{g}_c$ [95% CI]	0.091 **[0.035, 0.146]**	0.204 [-0.075, 0.482]
τ_c^2 [95% CI]	0.011 **[0.002, 0.019]**	0.028 [-0.070, 0.126]

Table 7. Two-class model of brief intervention duration moderation of use consequences effect sizes.

$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	
	Class 1	Class 2
Duration *Minutes (mean)*	21.2	108.6
k_c	97	6
(%)	(94.0)	(6.00)
k_{es}	38	6
$\bar{g}_c$ [95% CI]	0.116 **[0.073, 0.159]**	0.323 **[0.130, 0.516]**
τ_c^2 [95% CI]	0.011 **[0.002, 0.020]**	0.000 [-0.145, 0.146]

Intervention Components

Among the primary studies in the present analyses, the number and types of intervention components delivered during a BI session varied (see Table 1). Most BIs included prescriptive advice (62.8%), information booklets (56.2%), and/or normative feedback (72.7%), yet it is not directly apparent which combination of these or other components is most likely to be implemented. Results of a two-class mixture model of BI components are presented in Table 8, providing some insight into the configuration of components utilized across the 124 included studies. In the first class, composed of approximately 78 studies, the predominating component is prescriptive advice; decisional balance exercises, goal-setting activities, skills training, and referrals to services are unlikely to be used by studies in this class, and information booklets and normative feedback are approximately as likely to be used as not used. In the second class of about 43 studies, BIs are likely to include information booklets, decisional balance exercises, goal-setting activities, and normative feedback.

Table 8 also presents findings of the class-specific meta-analyses for the drug use outcome domain. In the first class, characterized by prescriptive advice, 26 studies provided effect sizes yielding an estimated BI effect of $\bar{g}_c$ = 0.085, 95% CI [0.033, 0.137], a reduction in drug use similar in magnitude and significance to the overall (single-class) estimate of $\bar{g}$ = 0.090, 95% [0.038, 0.143]. Between-study variation remained identical to the overall model, and maintained significance. In the second class, composed of five effect sizes from studies likely to use motivational components, the BI effect estimate was substantially larger, $\bar{g}_c$ = 0.206, 95% [-0.078, 0.491]. In this class, the estimate for between-study heterogeneity, τ_c^2 = 0.029, 95% CI [-0.080, 0.138], was

Table 8. Two-class model of intervention component moderation of drug use effect sizes.

$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	
	Class 1	Class 2
Advice		
1 (no)	0.27	0.55
2 (yes)	**0.73**	0.45
Booklet		
1 (no)	0.46	0.39
2 (yes)	0.54	**0.61**
Decisional Balance		
1 (no)	**0.89**	0.33
2 (yes)	0.11	**0.67**
Goal-Setting		
1 (no)	**0.73**	0.09
2 (yes)	0.27	**0.91**
Normative Feedback		
1 (no)	0.42	0.00
2 (yes)	0.58	**1.00**
Skills Training		
1 (no)	**0.87**	**0.84**
2 (yes)	0.13	0.16
Referral		
1 (no)	**0.84**	0.53
2 (yes)	0.16	0.47
k_c	78	43
(%)	(64.6)	(35.4)
k_{es}	26	5
$\bar{g}_c$ [95% CI]	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ_c^2 [95% CI]	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]

somewhat larger than in the advice class, τ_c^2 = 0.010, 95% CI [0.002, 0.018]. In the motivational class, both the effect size and the between-study variation estimates were not significantly different from zero. Meta-analyses of use consequences effect sizes

(Table 9) offer similar findings for the first, prescriptive advice component class. Here, summarizing the 41 available effect sizes leads to a BI effect and between-study variation estimate again similar to the overall (single-class) model. In the motivational component

Table 9. Two-class model of intervention component moderation of use consequences effect sizes.

$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	
	Class 1	Class 2
Advice		
1 (no)	0.27	0.55
2 (yes)	**0.73**	0.45
Booklet		
1 (no)	0.46	0.39
2 (yes)	0.54	**0.61**
Decisional Balance		
1 (no)	**0.89**	0.33
2 (yes)	0.11	**0.67**
Goal-Setting		
1 (no)	**0.73**	0.09
2 (yes)	0.27	**0.91**
Normative Feedback		
1 (no)	0.42	0.00
2 (yes)	0.58	**1.00**
Skills Training		
1 (no)	**0.87**	**0.84**
2 (yes)	0.13	0.16
Referral		
1 (no)	**0.84**	0.53
2 (yes)	0.16	0.47
k_c	78	43
(%)	(64.6)	(35.4)
k_{es}	41	7
$\bar{g}_c$ [95% CI]	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ_c^2 [95% CI]	0.014 **[0.005, 0.022]**	0.000 *

class (Class 2), seven available effect sizes lead to an estimate of BI effect that was both considerably larger than the advice class and significantly different from zero, $\bar{g}_c$ = 0.310, 95% CI [0.157, 0.464]. Between-study variation in this class was reduced to zero, but the standard error was inestimable.

Sample Characteristics

As described in Chapter II, characteristics of a primary study's sample, such as racial/ethnic or sex/gender composition or average age, may moderate a program's effectiveness. At the same time, it was noted that utilizing summary estimates of participant attributes in moderation analyses may induce aggregation or ecological bias, such that the between-study relation of the moderator with the program's effect differs from the within-study relation (in magnitude, direction, or both). Consequently, the utility of variables summarizing sample characteristics is limited in aggregate data meta-analysis. To investigate whether the present method, which in this case would examine the co-occurrence of several summary measures of participant attributes, may provide additional insight into the role of such characteristics in BI effectiveness, mixture models were fitted using each study's reported proportion of non-Hispanic white participants, proportion of female participants, and average participant age.

Table 10 presents the results of a two-class mixture model utilizing race/ethnicity, sex/gender, and age indicator variables. The first class (Class 1), composed of approximately 86 studies, was characterized by participant samples that were on average largely non-Hispanic white (74%) and male-identifying (female: 33%), with a mean age of 36.5 years. In the second class, 31 studies had a majority of participants who were racial/ethnic minority (non-Hispanic white: 24%) and female-identifying (56%), and had

a lower mean age (27.8 years). Findings of the meta-analyses of drug use effect sizes indicate that in the first class of studies, whose participants were majority older white individuals identifying as males, the BI effect estimate using 26 effect sizes was similar to the overall effect estimate and suggestive of a positive effect of BIs to reduce drug use, $\bar{g}_c$ = 0.085, 95% CI [0.033, 0.137]. In the second class, made up of studies with younger participants and with a larger proportion of participants with a racial/ethnic minority identity and who identified as female, synthesis of five effect sizes indicated a larger but nonsignificant BI effect, $\bar{g}_c$ = 0.206, 95% CI [-0.078, 0.491]. Between-study variation was not decreased from the overall (single-class) model in either class, and was somewhat increased in the second class.

Table 10. Two-class model of sample characteristic moderation of drug use effect sizes.

$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	
	Class 1	Class 2
Proportion Non-Hispanic White		
Mean	0.74	0.24
Proportion Female		
Mean	0.33	0.56
Average Age		
Mean (years)	36.5	27.8
k_c	86	31
(%)	(73.2)	(26.8)
k_{es}	26	5
$\bar{g}_c$ [95% CI]	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ_c^2 [95% CI]	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]

Results of within-class meta-analyses of use consequences effect sizes (Table 11) are similar to those of drug use effect sizes for the first class characterized by studies with participants who were older, non-Hispanic white, and male-identifying. In the second class, the BI effect was again substantially larger among study samples that had greater representation of younger racial/ethnic minority and female-identifying participants, $\bar{g}_c$ = 0.310, 95% CI [0.157, 0.464]; in contrast to the drug use model (Table 10), this larger BI effect maintained significance, and between-study heterogeneity was reduced to zero (the standard error for this quantity could not be estimated, however).

Table 11. Two-class model of sample characteristic moderation of use consequences effect sizes.

$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	
	Class 1	Class 2
Proportion Non-Hispanic White		
Mean	0.74	0.24
Proportion Female		
Mean	0.33	0.56
Average Age		
Mean (years)	36.5	27.8
k_c	86	31
(%)	(73.2)	(26.8)
k_{es}	41	7
$\bar{g}_c$ [95% CI]	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ_c^2 [95% CI]	0.014 **[0.005, 0.022]**	0.000 *

Table 12 shows an alternative, three-class model using the sample characteristic indicator variables. The first class is broadly similar to the first class of the two-class model (Table 10) in its composition (primarily studies with older non-Hispanic white individuals identifying as male) and in the findings of the within-class meta-analysis. The second class is characterized by studies with majority racial/ethnic minority participants (non-Hispanic white: 39%) and a substantial majority of female-identifying participants (96%), with a younger average age (34.9 years) compared to Class 1. The final class (Class 3) is composed of studies whose samples were majority non-Hispanic white (63%) and made up of an approximately equal proportion of male- and female-identifying participants on average (female: 52%). Participants in these studies were also substantially younger on average (25.4 years) compared to the first and second classes.

Table 12. Three-class model of sample characteristic moderation of drug use effect sizes.

$\bar{g}$ [95% CI]		0.090 **[0.038, 0.143]**	
τ^2 [95% CI]		0.010 **[0.001, 0.019]**	
	Class 1	Class 2	Class 3
Proportion Non-Hispanic White			
Mean	0.60	0.39	0.63
Proportion Female			
Mean	0.25	0.96	0.52
Average Age			
Mean (years)	38.8	34.9	25.4
k_c	69	39	9
(%)	(59.4)	(33.1)	(7.5)
k_{es}	23	5	3
$\bar{g}_c$ [95% CI]	0.077 [-0.029, 0.183]	0.174 [-0.339, 0.687]	0.333 [-0.140, 0.805]
τ_c^2 [95% CI]	0.010 [-0.001, 0.020]	0.000 [-0.052, 0.053]	0.059 [-0.126, 0.243]

In within-class meta-analyses of drug use effect sizes,[13] both the BI effect and between-study variation estimates were nonsignificant for all classes, likely because of reduced precision resulting from fewer effect sizes being available in most classes compared with the two-class model. Nevertheless, in comparison to the first class (composed of studies with primarily older non-Hispanic white participants identifying as male), the BI effect was larger in the class featuring studies with larger proportions of younger racial/ethnic minority participants identifying as female, $\bar{g}_c$ = 0.174. 95% [-0.339, 0.687]. Further, in the third class, composed of studies with participants who were on average the youngest, the largest BI effect was observed, $\bar{g}_c$ = 0.333, 95% CI [-0.140, 0.805]. Importantly, only five effect sizes were available to synthesize in the second class, while only three were available in the third class. Further, aside from Class 2, in which between-study variation was reduced to zero, estimates of between-study variation were either not decreased (Class 1) or were increased (Class 3) in comparison to the overall model.

[13] Class-specific meta-analyses of use consequences effect sizes for the three-class solution were underidentified, and are not presented.

CHAPTER V

CONCLUSIONS

Analyses presented in the previous chapter demonstrated a novel application of mixture modeling to characterize study, intervention, and sample-related drivers of effect size variation in meta-analysis. The aims of this application were to investigate whether and how these factors co-occurred in a sample of studies implementing a prevention program, and whether such co-occurrence would relate to or modify the program's effectiveness. Findings suggest the method meets both these aims. In the present section, findings will be discussed in view of their potential interpretive value and utility in future research on, and implementations of, brief interventions and other prevention programs. Further, identified limitations and modeling challenges, briefly noted in the prior chapter, will be considered in greater detail. The chapter will conclude with an outline of avenues for future research.

Table 13 summarizes meta-analytic findings presented in Chapter IV and additionally provides 95% prediction intervals for within-class meta-analyses. In the first models presented, aspects of primary studies' efficacy-to-effectiveness staging were examined. These attributes, which pertain to intervention flexibility, setting characteristics, provider type and training level, patient baseline severity and degree of incentivization, and degree of implementation support and monitoring, provide an indication of the feasibility, resource-intensiveness, and cost with which a program may be implemented. Clinically-typical or effectiveness-testing implementations may require fewer resources and by consequence have greater feasibility; at the same time, they may be at risk of implementation failure arising from poor adherence to an intervention's

Table 13. Summary of meta-analytic findings accompanied by within-class 95% prediction intervals (PI). Three-class sample characteristics model (Table 12) not shown.

		Within-class Models	
	Overall Model	*Class 1*	*Class 2*
	Efficacy-to-Effectiveness Staging – Drug Use		
k_c		80	39
k_{es}		26	5
$\bar{g}$ [95% CI] [a]	0.090 **[0.038, 0.143]**	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ^2 [95% CI] [b]	0.010 **[0.001, 0.019]**	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]
[95% PI]		[-0.129, 0.299]	[-0.507, 0.919]
	Efficacy-to-Effectiveness Staging – Use Consequences		
k_c		80	39
k_{es}		41	7
$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	0.014 **[0.005, 0.022]**	0.000 *
[95% PI]		[-0.145, 0.343]	[0.108, 0.512]
	Study Characteristics (Risks of Bias) – Drug Use		
k_c		76	48
k_{es}		26	5
$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]
[95% PI]		[-0.129, 0.299]	[-0.507, 0.919]
	Study Characteristics (Risks of Bias) – Use Consequences		
k_c		76	48
k_{es}		41	7
$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	0.014 **[0.005, 0.022]**	0.000 *
[95% PI]		[-0.145, 0.343]	[0.108, 0.512]
	Duration – Drug Use		
k_c		97	6
k_{es}		24	5
$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	0.091 **[0.035, 0.146]**	0.204 [-0.075, 0.482]
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	0.011 **[0.002, 0.019]**	0.028 [-0.070, 0.126]
[95% PI]		[-0.134, 0.316]	[-0.494, 0.902]

Table 13. (continued).

	Overall Model	Within-class Models *Class 1*	*Class 2*
		Duration – Use Consequences	
k_c		97	6
k_{es}		38	6
$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	0.116 **[0.073, 0.159]**	0.323 **[0.130, 0.516]**
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	0.011 **[0.002, 0.020]**	0.000 [-0.145, 0.146]
[95% PI]		[-0.101, 0.333]	[0.05, 0.596]
		Intervention Components – Drug Use	
k_c		78	43
k_{es}		26	5
$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]
[95% PI]		[-0.129, 0.299]	[-0.507, 0.919]
		Intervention Components – Use Consequences	
k_c		78	43
k_{es}		41	7
$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	0.014 **[0.005, 0.022]**	0.000 *
[95% PI]		[-0.145, 0.343]	[0.108, 0.512]
		Sample Characteristics – Drug Use	
k_c		86	31
k_{es}		26	5
$\bar{g}$ [95% CI]	0.090 **[0.038, 0.143]**	0.085 **[0.033, 0.137]**	0.206 [-0.078, 0.491]
τ^2 [95% CI]	0.010 **[0.001, 0.019]**	0.010 **[0.002, 0.018]**	0.029 [-0.080, 0.138]
[95% PI]		[-0.129, 0.299]	[-0.507, 0.919]
		Sample Characteristics – Use Consequences	
k_c		86	31
k_{es}		41	7
$\bar{g}$ [95% CI]	0.106 **[0.063, 0.150]**	0.099 **[0.054, 0.143]**	0.310 **[0.157, 0.464]**
τ^2 [95% CI]	0.014 **[0.005, 0.022]**	0.014 **[0.005, 0.022]**	0.000 *
[95% PI]		[-0.145, 0.343]	[0.108, 0.512]

[a] For within-class models, values correspond to $\bar{g}_c$. [b] For within-class models, values correspond to τ_c^2.

design or protocol (i.e., fidelity), inadequate provider training, or inadequate support for program delivery. Moreover, even when a program appears successfully implemented on measures such as provider uptake and patient satisfaction, it may be less effective, for instance, in the absence of sufficiently rigorous provider training. In research-typical or efficacy-testing implementations, by contrast, there may be greater resources for training, as well as support for intervention delivery (e.g., additional staff) and more intensive fidelity monitoring (e.g., direct observation and immediate corrective feedback). Intuitively, the presence of these characteristics could lead to greater program effect, yet they may be counterbalanced by other aspects of a research-typical implementation, such as having a provider that is contracted or in training. These providers may less effectively deliver the program because of a lack of clinical experience or existing relationship with the patient, both of which may be more common in clinically-typical implementations in which a provider is more likely to be a working clinician, including a patient's primary care provider. Similarly complex relations may be present for patients: clinically-typical patients who may present with a variety of severity and risk levels may also be generally more willing to accept the program, while at the same time, having fewer severe patients in the sample may attenuate the observed effect of the program. Conversely, among research-typical patients (who are more likely to exceed an elevated risk or severity threshold) there may be less acceptance of the program but simultaneously the possibility of observing a larger program effect given their more severe symptomatology.

The co-occurrence of efficacy-to-effectiveness factors and their relation to BI program effect was examined using a two-class mixture model. Classes were similar across several characteristics, including implementation among participants who were

more likely to have greater severity or risk, and/or to be incentivized for participation; use of briefer, less intensive BIs; and allowing providers moderate flexibility in BI delivery. Classes were distinguished, in particular, by their level of provider training, intervention support, and fidelity monitoring, as well as practitioner type and implementation context. Synthesis of effect sizes in the first class, which was characterized by a higher likelihood of intensive provider training and the provision of more extensive intervention support and monitoring, indicated that such studies have a comparatively small positive effect on reducing drug use and use consequences. Inspection of the prediction intervals for both outcome domains suggests that future implementations of similar studies would find effects that varied in both magnitude and direction.

Contrastingly, summarizing effect sizes in the second class of studies that had a higher probability of less intensive provider training, not providing extensive support or monitoring, and additionally being implemented in a community setting and with clinical providers, revealed an approximately three-fold larger BI effect compared to the research-typical class. Prediction intervals suggested that in future studies similar to those in the second class, BI effectiveness for drug use may range in both magnitude and direction, while for use consequences studies are likely to find a positive effect that may vary substantially in magnitude. Importantly, there were substantially fewer effect sizes available in the clinically-typical class than in the research-typical class, so this disparity in effect should be interpreted cautiously (this is the case for the smaller class of most models presented in Chapter IV). Despite this, these findings suggest that the research- or clinically-typical nature of a study's implementation may be associated with BI effect.

More specifically, each class markedly varied in its probability of featuring implementation monitoring, support, and intensive training, and yet the class characterized by the absence of such resources was found to have the greater overall BI effect. It may be the case that despite the evidenced relation of implementation fidelity with program effectiveness (Lipsey, 2009; Sanetti & Kratochwill, 2014; Sims & Melcher, 2017), the relative brevity and simplicity of BIs means they do not necessitate intensive monitoring, support, or substantial training to have some effect. Indeed, both classes were likely to utilize a shorter-duration and/or low-complexity BI.

Rather than extensive fidelity monitoring, support, and training, the key drivers of the larger BI effect in the clinically-typical class may be the confluence of implementation of the BI in a community or realistic setting, by practitioners or providers who have clinical experience and potentially a pre-existing relationship with the patient. Such a conclusion is supported by meta-analyses examining the role of providers in intervention effectiveness (Del Re et al., 2012) and the importance of the provider-patient alliance in achieving therapeutic outcomes (Martin et al., 2000), which conclude that positive alliance is consistently associated with effectiveness, but that this relation is moderated by the provider's ability to form and maintain alliance. The setting may also have some relation with the level of outcome severity (and responsiveness to intervention) among participants. It may be the case, then, that when BIs are delivered in community settings by providers whom patients trust or have comfort with, or who have experience in patient engagement, BIs have greater effect. In the context of the present method, which considered the co-occurrence of numerous potential efficacy-to-effectiveness factors, this conclusion also highlights a potential utility of the method in its

ability to leverage potentially counterintuitive findings related to some factors (e.g., fidelity monitoring and implementation support) to further probe for moderators.

The next series of models examined the co-occurrence of study characteristics that may increase biases or could otherwise influence program effectiveness. Many of these factors are codified in the Cochrane Risk of Bias Tool (Higgins et al., 2011), which is widely used in meta-analysis, while others were ad hoc measures that captured studies' missing data handling strategy, and whether implementation monitoring and problems were reported or identified. The results of a two-class mixture model indicated that studies in the motivating data set were likely to have high or unclear risk in the blinding of outcomes assessors (most studies used self-report measures, which are by definition unblinded), in the presence of selective reporting (most studies lacked registered protocols to assess selective reporting), and in the reporting modality (most studies used interview assessments rather than biological measure or self-administered assessments). Studies in both classes were also unlikely to report, or have coders identify, implementation problems.

Primary studies differed in their probability of risk associated with random sequence generation, allocation concealment, incomplete data, missing data handling, and presence or absence of intervention monitoring. In the first class, studies were more likely to have low risk of bias for most factors, which would reflect, for instance, computerized random sequence generation, fully concealed allocation to treatment or control groups, and the presence of monitoring. In the second class, studies were likely to have high or unclear risk across all differentiating factors, suggesting that studies in this class were more likely to be characterized by, for example, unclear randomization

strategy, poorly concealed allocations, substantial or unbalanced attrition, use of listwise deletion rather than a modern imputation approach, and an absence of monitoring. Taken together, then, the two classes could be interpreted as reflecting a lower overall risk of bias (the first class) or a higher overall risk of bias (the second class). In within-class meta-analyses, BIs implemented in studies in the class associated with a lower probability of risk of bias had a positive but comparatively small effect to reduce drug use and use consequences. In contrast, studies with greater probability of risk of bias had an approximately three-fold larger effect. As in the efficacy-to-effectiveness models, the latter effect estimate is derived from substantially fewer effect sizes than the former, and intervention monitoring (here measured as presence or absence not as degree or intensity) is again present in the class associated with lower BI effect. Prediction intervals for study characteristic models also have similar interpretations to the efficacy-to-effectiveness models: future studies with lower-risk profiles similar to those of studies in the first class would be expected to find effects that ranged in both magnitude and direction, while higher-risk studies may find effects on drug use that vary in magnitude and direction and effects on use consequences that are positive but range in magnitude.

A straightforward conclusion to draw from these findings is that the BI effect observed in the higher-risk class is upwardly biased, perhaps by poorly executed randomization and assignment to groups. Alternatively, differences in BI effect among these studies may be driven by high differential attrition, and by the same token, listwise deletion of participants missing follow-up data. Finally, the BI effect may be unrelated to these factors, and be larger for another reason (or set of reasons). Nonetheless, that a larger effect was seen in the higher-risk group across two distinct outcome domains (drug

use and use consequences) suggests some stability in the link between the study characteristics considered and the BI effect. In the context of BIs in particular, it may be that factors related to attrition and missing data handling are the overriding drivers of the larger overall effect observed. Supporting this conclusion is the observation that some alcohol-related BIs have greater effect among those with lower symptom severity (Baumann et al., 2018); at the same time, in longitudinal studies in general it is not uncommon for participants with greatest risk or severity to also be the most susceptible to loss to follow up (Ribisl et al., 1996). It is potentially the case, therefore, that studies in the group at high risk of incomplete data and bias-inflating missing data strategies were also more likely to have a pool of remaining participants who had less severe drug use or use consequences outcomes, among whom BIs have been shown to have greater effect.

The models presented in the prior chapter also examined aspects of the BI itself, beginning with the duration of the intervention. Duration, as a continuous variable reported in minute-units, was investigated for two reasons. The first arises from existing evidence on BI duration that suggests duration does not moderate BI effectiveness (Beyer et al., 2018; Kaner et al., 2018). Importantly, in these studies meta-regression was used to assess whether duration was an important moderator, and by consequence minute-unit-change in duration was the predictor of interest. Single-minute increases in BI duration, intuitively, may not measurably impact effectiveness. Thus, as noted above, it may be more informative to examine whether studies may be grouped around an average BI duration, and to investigate this categorical difference in average BI duration for its influence on BI effect. Such an analysis was presented in the form of a two-class mixture model, which found that most BIs were, on average, 21.2 minutes in length. A second

group of studies was considerably longer in average duration (108.6 minutes), an amount that could be accumulated over several BI sessions. Shorter-duration BIs were found to have a comparatively small overall effect on drug use and use consequences that would be expected to vary in both direction and magnitude in future implementations, while the longer-duration interventions had a substantially larger effect, which may range in magnitude and direction for future studies of drug use but would likely remain positive for use consequences. These findings suggest BI duration may be linked to effectiveness when broader categorical rather than incremental differences in duration are considered. Importantly, a benefit of mixture modeling in this scenario is that use of arbitrary cut points to create duration groups was unnecessary, and instead such groups could be derived empirically. Indeed, other studies that examined the moderation role of BI duration (Black et al., 2016; Tanner-Smith & Lipsey, 2015) used apparently arbitrarily-defined categories to represent duration (e.g., less than or equal to 30 minutes, or greater than 30 minutes), and found no link between duration and BI effect. This contrasting finding underlines the potential utility of empirical cut points, which can be readily extended to other prevention programming where there are open questions about the optimal duration of intervention exposure or dosage.

In addition to duration, the number and types of intervention components can vary from one implementation of a BI to another. The variety of components that can be incorporated into BIs is, on the one hand, a strength of the approach in that BIs can be highly tailored to patient severity or risk level, readiness to change, and receptivity to or preference for certain intervention activities. On the other hand, the lack of standardization in BI structure and composition has complicated determining BI

effectiveness, given that the individual implementations vary considerably in the components used, and in whether those components are required to be administered to all patients or are selected on a case-by-case basis using provider discretion. Relatedly, it has been challenging to identify whether some components are essential to program effectiveness, i.e., represent indispensable intervention *kernels* that are key to BI success (Embry & Biglan, 2008; Tanner-Smith & Lipsey, 2015).

Despite inconsistencies in components used, however, an effort to broadly categorize the types of components used in BIs can begin with a consideration of whether the BI is underlain by a motivational or prescriptive theoretical stance. The former, pioneered by Miller (1991, 2002), is oriented toward strengthening patient self-efficacy and motivation for change, while the latter is driven more by practical considerations such as existing clinician comfort with traditional delivery of authoritative advice and the brevity with which such advice can be proffered (Van Voorhees et al., 2009). With this line of thinking in mind, a two-class mixture model of common BI components was presented. Model results indicated that the majority of studies had a high probability of utilizing prescriptive advice, and were unlikely to use decisional balance, goal-setting, or skills training components, or to make referrals to other services. Conversely, the remaining studies were likely to make use of decisional balance and goal-setting components, and to provide a take-home information booklet and personalized normative feedback. Thus, studies aligned with the characterization of prescriptive (the first class) or motivational (the second class) in orientation.

Meta-analyses carried out within classes found that studies likely to be prescriptive (advice-based) in nature had a comparatively small and positive overall

effect to reduce drug use and use consequences (that may range in both magnitude and direction in future implementations), while BIs more likely to be motivational had a considerably larger positive effect that would be expected to remain positive for use consequences but may vary in magnitude and direction for drug use. Because in the present analyses, the co-occurrence of components was modeled, findings are suggestive of prescriptive advice serving as a modestly effective kernel, in that when implemented as a standalone or primary BI component, it does produce some positive effect. By comparison, however, decisional balance, goal-setting, normative feedback, and information booklet components, when implemented (or available to implement) together, yielded a much larger BI effect. Such a finding may indicate that these components are kernels in and of themselves, an interpretation that has been examined using meta-regression (Tanner-Smith & Lipsey, 2015). Perhaps more critically, it suggests that there is an important interrelationship of these components, one that may be tied to their common focus on enhancing patient self-efficacy to alter their behavior. The same interrelationship may also evidence the complexity and multidimensionality of behavior change, and indicate that components targeting multiple sources of motivation (e.g., awareness of risk information and norms, or prospective goal-setting) and stages of change (e.g., contemplation of pros and cons of use) are necessary to modify behavior (Dempsey et al., 2018; Perkins & Berkowitz, 1986; Prochaska & DiClemente, 1984). Considering the co-occurrence of these components and the larger BI effect observed together, findings of these models lend support to the use of motivational rather than prescriptive BI components.

The final factors examined as potential drivers of effect size variation were sample or participant characteristics, specifically racial/ethnic and sex/gender composition, and average participant age. As noted in Chapter II, use of measures that summarize participant characteristics (i.e., are participant-variant within studies) can produce an aggregation bias that misrepresents the moderation effect of interest. Optimally, such moderation relations are explored using individual participant data meta-analysis, in which the relations are examined within-study and summarized. This approach preserves the within-study nature of the relation, while also maintaining the aim of meta-analysis to synthesize the findings of numerous primary studies (and, here, the study-level influence of a moderator on those findings). Crucially, individual participant data meta-analyses are resource-intensive and as yet rarely undertaken in comparison to aggregate data meta-analysis.

To investigate whether the present method may offer some value for examining the role of participant characteristics in BI effectiveness in the absence of individual participant data, two- and three-class models were estimated using three continuous summary measures of primary study sample attributes: proportion non-Hispanic white, proportion female, and average participant age. In the two-class solution, the first class was characterized by studies with a higher proportion of non-Hispanic white individuals identifying as male, who were comparatively older than participants in the second class, in which studies also had higher proportions of racial/ethnic minority and female-identifying participants. The three-class solution effectively separated the latter class into two classes: one primarily composed of studies with somewhat younger, racial/ethnic minority female-identifying participants, and another characterized by studies with

majority non-Hispanic white participants approximately balanced across sexes/genders, but having samples with the youngest average age of all classes. All class-specific meta-analyses found a positive BI effect for reducing drug use and use consequences; the smallest effect was found for the classes predominated by studies with older (mean ages of 36.5–38.8), non-Hispanic white individuals identifying as male, while the largest effect was observed for classes having studies with a higher proportion of participants who were younger (mean age of 25.4–27.8), had a racial/ethnic minority identity, and who identified as female. As with prior models, most effects would be expected to vary in both magnitude and direction in future implementations, with the exception of the studies with a similar sample composition to those in the second class of the two-class solution (i.e., studies with younger participants and with higher proportions of racial/ethnic minority and female-identifying participants); here, BI effects on use consequences are likely to remain positive but could vary in magnitude.

When the findings of the two- and three-class models are viewed together, they suggest that participant race/ethnicity and sex/gender may have some role in BI effectiveness, and that participant age may also moderate effectiveness, given the consistently larger effect size seen as average age decreased across classes, while racial/ethnic and sex/gender composition varied across the same classes. This conclusion is supported by evidence showing that BIs are effective for young adults ages 18–30 (Fachini et al., 2012; Tanner-Smith & Lipsey, 2015), but have unclear or no effectiveness for older adults (Fleming et al., 1999; Monti et al., 1999). Older adults may be more likely to have experienced chronic use and have more intractable use behaviors, and the ongoing nature of their use paired with the relatively low treatment intensity

characteristic of BIs compared to other treatment modalities (e.g., ongoing therapeutic or in-patient substance use treatment) may render BIs less effective for this demographic (see Saitz, 2010). The interpretation that foregrounds age (and to a lesser extent, female sex/gender) as the factor driving BI effectiveness is in light of several studies that have found BIs to be ineffective among racial and ethnic minorities, perhaps as a consequence of poor (or entirely absent) cultural adaptation (see Manuel et al., 2015), and to be generally effective among individuals identifying as female (Ballesteros et al., 2004; Manwell et al., 2000; O'Connor & Whaley, 2007; Ondersma et al., 2007). Finally, as was noted for models examining BI duration, a benefit of the present method for examining continuously-measured participant characteristics is that decisional cut points were unnecessary; instead, groupings of characteristics were defined empirically.

The above interpretations underline the key feature of the present method: rather than examining the influence of a potential effect moderator in isolation, the co-occurrence of multiple moderators is modeled. In this fashion, moderators are investigated in a way that is more reflective of realistic program implementation, in which numerous implementation-related factors may work in concert to alter the effectiveness of a program. In the present analyses, a finding in which this utility is particularly apparent is related to fidelity monitoring. Despite the strong prior evidence that higher fidelity is associated with greater program effectiveness, when considered alongside other implementation and methodological factors (e.g., implementation setting, provider characteristics, attrition, and missing data handling), the importance of fidelity monitoring appears to be diminished. This conclusion is not an argument *against* fidelity monitoring, but instead suggests that for a brief, low-intensity intervention such as BIs,

greater effect may be achieved by focusing resources on the fit and installation of the intervention in appropriate community settings, rather than on intensive fidelity monitoring.

More broadly, the method described here merges two analytic frameworks that are useful for exploring rich and complex data: finite mixture modeling and meta-analysis. The benefit of utilizing mixture modeling to investigate meta-analytic data related to prevention program effectiveness is that such data is inherently complex: composed of multiple independent trials, each assessing a program's effectiveness using a variety of designs, with different levels of methodological rigor, in distinct settings, and among potentially dissimilar participant populations. Further, as is the case with BIs, the program itself may differ in composition, theoretical underpinning, delivery modality, and duration – a reality that adds both informative and nuisance variation to the data. In view of such complexities, mixture modeling can be used to parse relations among numerous factors that potentially drive effect size variation into useable representations of their co-occurrence in practice. These representations – or implementation profiles – can then serve as the basis for synthesizing effects found by studies with such profiles to determine whether the interrelationship among moderators drives variation in program effectiveness, as demonstrated here.

Limitations and Considerations

The first and most prominent limitation of the present method is the number of studies needed for stable mixture model estimation. While there are no firm guidelines for minimal sample size in mixture modeling, recommendations have ranged into the several-hundreds of observations (e.g., Nylund et al., 2007). Nevertheless, other factors

play into estimability of mixture models, particularly number of classes estimated, class separation, the true or underlying class structure, and number and types of indicator variables (Gudicha et al., 2016; Masyn, 2013). When classes are well separated, for example, fewer observations (here, studies) are needed to derive useable parameter estimates. Similarly, continuous indicator variables (e.g., program duration) contribute comparatively more information to model estimation than do categorical variables, and a model with one or multiple continuous indicators may require fewer studies. In its current form, then, the present method may be best suited to larger meta-analyses; nevertheless, in more moderately-sized meta-analyses, estimation of exploratory models can be used to assess model convergence and stability, as well as interpretability arising from quality of class separation. When models indicate low entropy, when classes are poorly distinguished (i.e., most response probabilities are near 0.5), or when information criteria and other fit statistics suggest a multiple-class structure is unlikely, more traditional moderation analysis methods such as meta-regression or subgroup analysis could be used.

Another quantity that, when limited, produces analytic challenges is the number of effect sizes available within a class. It is to be expected in meta-analyses of prevention programs that not all studies will provide effect sizes for all outcomes, yet when there is substantial imbalance in the number of effect sizes within classes, estimation of meta-analytic models becomes challenging.[14] Especially impacted is estimation of the within-class between-study variation parameter, τ_c^2. It is known that when few effect sizes are

[14] For example, exploratory models using effect sizes from the alcohol use outcome domain found that one class routinely had only a single effect size available, preventing the estimation of within-class meta-analyses.

available, estimates of τ^2 can be downwardly biased with the use of unrestricted maximum likelihood estimation (Viechtbauer, 2005), which is utilized here given the use of mixture modeling. Thus, in models presented in the previous chapter, it is possible that in classes with a low number of effect sizes available, within-class between-study variation estimates may be larger than suggested by the reported τ_c^2. One strategy that may go some way to improving accuracy and reducing biasedness of the between-study variation estimate is the use of prior information. In this approach, predictive distributions of between-study variation derived from prior meta-analyses of BIs or similar prevention programs would be used as mildly informative priors for within-class meta-analyses (Turner et al., 2012). Predictive distributions have been estimated for τ^2 in the context of clinical interventions (e.g., Rhodes et al., 2015; Turner et al., 2015), but there is a need to develop similar estimates for prevention programming. Nevertheless, this is a growing area of research that could be leveraged in the present method.

A related issue is that in all models, substantial differences in BI effect between classes were found; in many instances, however, the within-class estimate of between-study variation was not reduced from the overall (single-class) model. In the cases when the class-specific between-study variation estimate was similar or unchanged from the overall model, the majority of studies had been assigned to that class; consequently, it is likely that class contained the same studies that contributed most between-study variation in the overall model. Thus, the magnitude of variation in the largest class was in general similar to the overall model. In the smaller classes (which typically had the larger BI effect estimate), it is possible that the studies assigned were those that, on average, had larger effect sizes, but that those effect sizes had greater variation. This conclusion

accords with the relatively low number of available effect sizes in those same classes: few effect sizes, of comparatively large but dissimilar values, would appear as having increased between-study variation. It may be that a remedy for this and other issues that occur when few effect sizes are available is the use of imputation to recover missing or unreported effect size information. Multiple imputation of meta-analytic data is challenging, however, given the complex structure of the data, the use of summary measures (e.g., mean group values), and practicalities involved with pooling multiple data sets. Similarly, issues are posed in mixture modeling due to the typical distributional assumptions of multiple imputation, which can obscure the presence of component distributions in the imputed data (Sterba, 2016). Nevertheless, imputation in meta-analytic contexts is an area of research with recent developments (including the use of pattern mixture modeling; Mavridis & White, 2020) that may be utilized in future implementations of the present method.

A final consideration related to effect size availability is that when there are few effect sizes available to synthesize, within-class statistical power and precision are reduced compared to the overall analysis using all available effect sizes. Together, these conditions limit the value of assessments of statistical significance of program effect or between-study variation estimates, especially when nonsignificance is indicated. As such, when there are a limited number of effect sizes available in any class, it may be most appropriate to consider only the effect estimate, and to do so while recognizing that the proximity of that value to the true population value is not known. Alternatively, or perhaps in addition to limiting interpretation to the effect estimates, Bayes factors (e.g., Dienes, 2014) could potentially be used to investigate whether within-class findings

accord with existing evidence on the association of certain moderators with program effectiveness (i.e., whether a null finding represents a rejection of that evidence or a model-specific sensitivity issue).

With regard to moderator variable selection, a potential limitation arises when there is interest in modeling continuous and categorical indicators simultaneously. Because continuous variables are typically considerably more informative than categorical variables (that is, each observation can be unique, whereas categorical values repeat across potentially numerous observations), the underlying component distributions identified in the model may be heavily influenced by the comparatively greater information provided by the continuous indicator(s). When this occurs, the categorical indicators may appear to poorly distinguish between classes because the class structure is more strongly driven by the continuous indicator(s). For instance, during preliminary model fitting, a model (not shown) was estimated that included BI duration as well as categorical indicators for provider and setting type. In this case, classes were identified that were nearly identical to those presented in Tables 6 and 7 with respect to fit, average duration in each class, and studies grouped into each class. At the same time, levels of the categorical provider and setting variables exhibited close to chance probabilities of occurring in each class, suggesting that the classes were primarily component distributions of BI duration rather than distributions reflecting the co-occurrence of certain types of providers and settings alongside differing average durations (the optimal interpretation of such a model). This consideration underlines the importance of model interpretability as a criterion for model selection. Indeed, despite the likely interrelation of duration and provider and setting types that may motivate an interest in modeling these

characteristics together, the nature of the variables themselves limits the interpretability of the model. One strategy that may go some way to addressing this issue is to first estimate a mixture model of the continuous variable of interest, from which approximate category levels could be empirically derived. This categorical variable could then be included in a mixture with other categorical indicators of interest, an approach which maintains some information from a continuous indicator but may limit the degree to which it compromises the interpretability of categorical indicators.

Lastly, in the final set of models that examined participant characteristics, the degree to which aggregation or ecological bias influenced the findings could not be assessed because individual participant data were not available. It will be important to examine the role of aggregation bias in such analyses using simulation studies or empirical aggregate data that has accompanying individual participant data, but in the interim it should be noted that the analyses presented here do not necessarily aim to provide findings similar to regression-based investigations of effect moderation by, for instance, mean age or proportion race/ethnicity or sex/gender. That is, what is of interest is not the linear relation of participant characteristics with BI effect – a quantity whose interpretation can be distorted by aggregation bias – but instead categorical profiles of studies represented by average values of the characteristics.

Indeed, the estimation of such profiles (classes) *independent* of effects (i.e., in a stage prior to moderation analysis), is an important advantage of this method. Specifically, when the association of the classes with effects is assessed, the relation that is examined is between the effect estimates and values of participant characteristics that are to some degree representative of several studies (such as the means of the indicators

from all studies assigned to a class). This is in contrast to meta-regression, in which the relation of each study's value of the characteristic (e.g., mean age) with the effect estimate is assessed. In real terms, should mean age be found to moderate effects in a meta-regression model, differences in program effect would be anticipated in a study with older participants compared to a study with younger participants, regardless of the nature of the within-study relation between age and program effect in those two studies. By contrast, in the present method the interest is in the relation of program effect with *groups* of numerous studies having, on average, older participants or younger participants. This qualitative or categorical difference contains information about the relation of age with program effects drawn from many studies, and when combined with measures of additional participant characteristics, the categorical profiles that are generated represent far more information than a summary statistic of a single study for a single moderator (given that values of other characteristics must be held constant when using meta-regression). As a result, such profiles may have potential to enhance the external validity or generalizability of prevention programs by providing insight on who can benefit most from the program, across multiple characteristics that may include participant age, racial/ethnic or sex/gender identity, or other attributes. This knowledge could be used to tailor program recruitment or outreach, or inform decisions about installation setting and populations of focus in support of more pragmatic program implementations (i.e., that are more realistic and relevant to contexts and participants; Glasgow, 2013).

The stagewise approach that separates modeling the co-occurrence of potential moderator variables from modeling moderation also differentiates the method from prior

applications of mixture modeling in meta-analysis. In those cases, effect sizes themselves were analyzed using mixture modeling, in an effort to identify homogeneous subgroupings of effects but not to explain why those subgroups might occur. Here, the subject of the mixture models are study characteristics, and insight into their co-occurrence can provide useful information independent of whether that co-occurrence meaningfully moderates program effectiveness. Put another way, should no difference in class-specific program effects be discovered, the method still provides an understanding of to what degree – and at what frequency – numerous aspects of a program's implementation co-occur in practice.

Future Research

In addition to investigating the use of stabilizing prior information to improve estimates of within-class between-study variation magnitude, noted in the previous section, future research will examine the compatibility of the illustrated method with a routine scenario in meta-analyses of prevention program effects: the availability of multiple, dependent effect sizes from primary studies. Effect size dependency occurs when studies collect multiple measures of the outcome of interest or collect data over several time points. A benefit of the present method is that once classes are formed with study-level indicator variables, the generated class assignment weights apply to the study overall; thus, these weights may then be associated with multiple effect sizes so that they are analyzed within the appropriate class. Importantly, however, standard random-effects models for meta-analysis as were used here do not preserve and properly model dependency in effect sizes, and instead, multivariate or three-level meta-analyses will need to be fitted within classes. Carrying out this modeling approach will be the next

stage of research for the present method. A further step will be to conduct simulation studies to examine whether use of the method with summary measures of participant attributes provides reasonably unbiased information on the relation of these characteristics with program effects, in lieu of individual participant data. Future applications of the approach will also explore the use of bootstrapped likelihood ratio tests for mixture model selection (Nylund et al., 2007).

Should the above research be fruitful, a longer-term goal is to broaden access to the approach through the development of an open-source implementation, such as for the R statistical environment. A second long-term aim is to explore full latent regression in the context of meta-analysis, which would augment the present method by also fitting mixture models directly to effect size data (e.g., Böhning, 2005; Schlattmann, 2009; van Houwelingen et al., 2002; Xia et al., 2005). Moderator classes would then be assessed for their relation with latent groupings of effects. A related aim is to fit all moderators simultaneously rather than as separate models as illustrated here. In this application, a higher-order latent variable would capture interrelations among lower-order latent variables representing individual groups of moderators (e.g., those pertaining to risks of bias, efficacy-to-effectiveness, etc.).

In summary, the method demonstrated here offers a novel application and integration of mixture modeling and meta-analysis to characterize drivers of between-study variation in prevention program effects. The method adds to existing approaches for identifying and investigating moderators of program effectiveness, such as meta-regression and subgroup analysis, by providing a framework to examine the co-occurrence of potential moderators in a way that parallels their interrelatedness in the

naturalistic implementation of a program. As a result, the method has the potential to strengthen the installation, implementation, and maintenance of prevention programs by providing insight into which methodological, intervention, and participant characteristics – alone or in combination – are critical for program effectiveness.

APPENDIX

MODEL FIT INFORMATION

Number of classes	BIC [a]	Entropy	aLRT [b] p-value
	Efficacy-to-Effectiveness		
1	1234.84	–	–
2	1173.15	0.89	0.00
3	1193.89	0.81	0.01
	Study Characteristics (Risks of Bias)		
1	1487.55	–	–
2	1498.83	0.59	0.24
3	1529.60	0.84	0.09
	Intervention Duration		
1	967.17	–	–
2	906.70	0.99	0.00
3	908.21	0.81	0.55
	Intervention Components		
1	1057.71	–	–
2	1046.62	0.68	0.01
3	1060.92	0.75	0.01
	Sample Characteristics		
1	924.94	–	–
2	916.18	0.69	0.04
3	911.58	0.81	0.15
4	911.63	0.73	0.15

[a] Bayesian information criterion. [b] Adjusted likelihood ratio test.

REFERENCES CITED

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Asparouhov, T., & Muthén, B. O. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181

Asparouhov, T., & Muthén, B. O. (2018). *Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model* (No. 21; Mplus Web Notes). https://www.statmodel.com/download/asparouhov_muthen_2014.pdf

Baker, W. L., White, C. M., Cappelleri, J. C., Kluger, J., & Coleman, C. I. (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice*, *63*(10), 1426–1434. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*(04), 520–540. https://doi.org/10.1093/pan/mpu003

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289. https://doi.org/10.1080/10705511.2015.1049698

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311. https://doi.org/10.1177/0081175012470644

Bakk, Z., & Vermunt, J. K. (2015). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31. https://doi.org/10.1080/10705511.2014.955104

Ballesteros, J., González-Pinto, A., Querejeta, I., & Ariño, J. (2004). Brief interventions for hazardous drinkers delivered in primary care are equally effective in men and women. *Addiction*, *99*(1), 103–108. https://doi.org/10.1111/j.1360-0443.2004.00499.x

Baumann, S., Gaertner, B., Haberecht, K., Bischof, G., John, U., & Freyer-Adam, J. (2018). How alcohol use problem severity affects the outcome of brief intervention delivered in-person versus through computer-generated feedback letters. *Drug and Alcohol Dependence*, *183*, 82–88. https://doi.org/10.1016/j.drugalcdep.2017.10.032

Berlin, J. A., & Antman, E. M. (1992). Advantages and limitations of meta-analytic regressions of clinical trials data. *Controlled Clinical Trials*, *13*(5), 422. https://doi.org/10.1016/0197-2456(92)90151-O

Beyer, F., Lynch, E., & Kaner, E. (2018). Brief interventions in primary care: An evidence overview of practitioner and digital intervention programmes. *Current Addiction Reports*, *5*(2), 265–273. https://doi.org/10.1007/s40429-018-0198-7

Black, N., Mullan, B., & Sharpe, L. (2016). Computer-delivered interventions for reducing alcohol consumption: Meta-analysis and meta-regression using behaviour change techniques and theory. *Health Psychology Review*, *10*(3), 341–357. https://doi.org/10.1080/17437199.2016.1168268

Böhning, D. (2005). Meta-analysis: A unifying meta-likelihood approach framing unobserved heterogeneity, study covariates, publication bias, and study quality. *Methods of Information in Medicine*, *44*(01), 127–135. https://doi.org/10.1055/s-0038-1633931

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27. https://doi.org/10.1093/pan/mph001

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*, 134–143. https://doi.org/10.1007/s11121-013-0377-7

Cheung, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods*, *13*(3), 182–202. https://doi.org/10.1037/a0013163

Cheung, M. W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 429–454. https://doi.org/10.1080/10705511.2013.797827

Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118957813

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social behavioral, and health sciences*. Wiley.

Davis, M. F., Shapiro, D., Windsor, R., Whalen, P., Rhode, R., Miller, H. S., & Sechrest, L. (2011). Motivational interviewing versus prescriptive advice for smokers who are not ready to quit. *Patient Education and Counseling*, *83*(1), 129–133. https://doi.org/10.1016/j.pec.2010.04.024

Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012). Therapist effects in the therapeutic alliance–outcome relationship: A restricted-maximum likelihood meta-analysis. *Clinical Psychology Review*, *32*(7), 642–649. https://doi.org/10.1016/j.cpr.2012.07.002

Dempsey, R. C., McAlaney, J., & Bewick, B. M. (2018). A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. *Frontiers in Psychology*, *9*, 1–16. https://doi.org/10.3389/fpsyg.2018.02180

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review*, *11*(3), 75–113. https://doi.org/10.1007/s10567-008-0036-x

Eusebi, P., Reitsma, J. B., & Vermunt, J. K. (2014). Latent class bivariate model for the meta-analysis of diagnostic test accuracy studies. *BMC Medical Research Methodology*, *14*(1), 88. https://doi.org/10.1186/1471-2288-14-88

Fachini, A., Aliane, P. P., Martinez, E. Z., & Furtado, E. F. (2012). Efficacy of brief alcohol screening intervention for college students (BASICS): A meta-analysis of randomized controlled trials. *Substance Abuse Treatment, Prevention, and Policy*, *7*(1), 40. https://doi.org/10.1186/1747-597X-7-40

Fleming, M. F., Manwell, L. B., Barry, K. L., Adams, W., & Stauffacher, E. A. (1999). Brief physician advice for alcohol problems in older adults: A randomized community-based trial. *The Journal of Family Practice*, *48*(5), 378–384.

Glasgow, R. E. (2013). What does It mean to be pragmatic? Pragmatic methods, measures, and models to facilitate research translation. *Health Education & Behavior*, *40*(3), 257–265. https://doi.org/10.1177/1090198113486805

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231. https://doi.org/10.1093/biomet/61.2.215

Gudicha, D. W., Tekle, F. B., & Vermunt, J. K. (2016). Power and sample size computation for Wald tests in latent class models. *Journal of Classification*, *33*(1), 30–51. https://doi.org/10.1007/s00357-016-9199-1

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

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

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. https://doi.org/ 10.1002/sim.1186

Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer.

Kaner, E. F., Beyer, F. R., Muirhead, C., Campbell, F., Pienaar, E. D., Bertholet, N., Daeppen, J. B., Saunders, J. B., & Burnand, B. (2018). Effectiveness of brief alcohol interventions in primary care populations. *The Cochrane Database of Systematic Reviews*, *2018*(2). https://doi.org/10.1002/14651858.CD004148.pub4

Kaner, E. F., Campbell, C., Pienaar, E. D., Heather, N., Schlesinger, C., & Saunders, J. (2003). Brief interventions for excessive drinkers in primary care health settings [Study protocol]. *The Cochrane Database of Systematic Reviews*, *2*. https://doi.org/10.1002/14651858.CD004148

Kypri, K., Langley, J. D., Saunders, J. B., Cashell-Smith, M. L., & Herbison, P. (2008). Randomized controlled trial of web-based alcohol screening and brief intervention in primary care. *Archives of Internal Medicine*, *168*(5), 530–536. https://doi.org/10.1001/archinternmed.2007.109

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling*, *20*(1), 1–26.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, *4*(2), 124–147. https://doi.org/10.1080/15564880802612573

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Manuel, J. K., Satre, D. D., Tsoh, J., Moreno-John, G., Ramos, J. S., McCance-Katz, E. F., & Satterfield, J. M. (2015). Adapting screening, brief intervention and referral to treatment (SBIRT) for alcohol and drugs to culturally diverse clinical populations. *Journal of Addiction Medicine*, *9*(5), 343–351. https://doi.org/10.1097/ADM.0000000000000150

Manwell, L. B., Fleming, M. F., Mundt, M. P., Stauffacher, E. A., & Barry, K. L. (2000). Treatment of problem alcohol use in women of childbearing age: Results of a brief intervention trial. *Alcoholism: Clinical and Experimental Research*, *24*(10), 1517–1524. https://doi.org/10.1111/j.1530-0277.2000.tb04570.x

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*(3), 438–450.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 551–611). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0025

Mavridis, D., & White, I. R. (2020). Dealing with missing outcome data in meta-analysis. *Research Synthesis Methods*, *11*(1), 2–13. https://doi.org/10.1002/jrsm.1349

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, *6*, 355–378. https://doi.org/10.1146/annurev-statistics031017-100325

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons, Inc.

Miller, W. R., & Rollnick, S. (1991). *Motivational interviewing: Preparing people to change addictive behavior*. The Guilford Press.

Miller, W. R., & Rollnick, S. (2002). *Motivational interviewing: Preparing people for change* (2nd ed.). The Guilford Press.

Monti, P. M., Colby, S. M., Barnett, N. P., Spirito, A., Rohsenow, D. J., Myers, M., Woolard, R., & Lewander, W. (1999). Brief intervention for harm reduction with alcohol-positive older adolescents in a hospital emergency department. *Journal of Consulting and Clinical Psychology*, *67*(6), 989–994. https://doi.org/10.1037/0022-006X.67.6.989

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling*. Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (2019). *Mplus* (Version 8.4) [Computer software]. Muthén & Muthén.

Nelson, L. R., & Zaichkowsky, L. D. (1979). A case for using multiple regression instead of ANOVA in educational research. *The Journal of Experimental Education*, *47*(4), 324–330. https://doi.org/10.1080/00220973.1979.11011701

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

O'Connor, M. J., & Whaley, S. E. (2007). Brief intervention for alcohol use by pregnant women. *American Journal of Public Health*, *97*(2), 252–258. https://doi.org/10.2105/AJPH.2005.077222

Ondersma, S. J., Svikis, D. S., & Schuster, C. R. (2007). Computer-based brief intervention: A randomized trial with postpartum women. *American Journal of Preventive Medicine*, *32*(3), 231–238. https://doi.org/10.1016/j.amepre.2006.11.003

Parr, N. J., Schweer-Collins, M. L., Darlington, T. M., & Tanner-Smith, E. E. (2019). Meta-analytic approaches for examining complexity and heterogeneity in studies of adolescent development. *Journal of Adolescence*, *77*, 168–178. https://doi.org/10.1016/j.adolescence.2019.10.009

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International Journal of the Addictions*, *21*(9–10), 961–976. https://doi.org/10.3109/10826088609077249

Prochaska, J. O., & DiClemente, C. C. (1984). *The transtheoretical approach: Crossing traditional boundaries of therapy*. Dow Jones-Irwin.

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 3.6.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rhodes, K. M., Turner, R. M., & Higgins, J. P. T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, *68*, 52–60.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*(1), 1–25. https://doi.org/10.1016/0149-7189(95)00037-2

RStudio Team. (2019). *RStudio: Integrated development environment for R* (Version 1.2.1335) [Computer software]. RStudio, Inc. http://www.rstudio.com/

Saitz, R. (2010). Alcohol screening and brief intervention in primary care: Absence of evidence for efficacy in people with dependence or very heavy drinking. *Drug and Alcohol Review*, *29*(6), 631–640. https://doi.org/10.1111/j.1465-3362.2010.00217.x

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

Sanetti, L. M. H., & Kratochwill, T. R. (Eds.). (2014). *Treatment integrity: A foundation for evidence-based practice in applied psychology*. American Psychological Association Press (Division 16).

Schlattmann, P. (2009). Investigating and analyzing heterogeneity in meta-analysis. In *Medical Applications of Finite Mixture Models* (pp. 153–199). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68651-4_7

Schlattmann, P., Verba, M., Dewey, M., & Walther, M. (2015). Mixture models in diagnostic meta-analyses—Clustering summary receiver operating characteristic curves accounted for heterogeneity and correlation. *Journal of Clinical Epidemiology*, *68*(1), 61–72. https://doi.org/10.1016/j.jclinepi.2014.08.013

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sims, B., & Melcher, B. (2017). Active implementation frameworks: Their importance to implementing and sustaining effective mental health programs in rural schools. In K. D. Michael & J. P. Jameson (Eds.), *Handbook of Rural School Mental Health* (pp. 339–361). Springer International Publishing. https://doi.org/10.1007/978-3-319-64735-7_22

Sterba, S. K. (2016). Cautions on the use of multiple imputation when selecting between latent categorical versus continuous models for psychological constructs. *Journal of Clinical Child & Adolescent Psychology*, *45*(2), 167–175. https://doi.org/10.1080/15374416.2014.958839

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis. *Journal of Substance Abuse Treatment*, *51*, 1–18.

Tanner-Smith, E. E., Saitz, R., Gelberg, L., Darlington, T., Parr, N. J., Schweer-Collins, M., & Frankel, L. (2020). *Brief substance use interventions in general healthcare settings meta-analysis*. https://osf.io/m48g6/

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, *309*(6965), 1351–1355. https://doi.org/10.1136/bmj.309.6965.1351

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*, 1559–1573. https://doi.org/10.1002/sim.1187

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338

Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., & Higgins, J. P. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*, *41*(3), 818–827. https://doi.org/10.1093/ije/dys041

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*, *34*(6), 984–998. https://doi.org/10.1002/sim.6381

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*(4), 589–624. https://doi.org/10.1002/sim.1040

Van Lissa, C. J. (2017). *MetaForest: Exploring heterogeneity in meta-analysis using random forests* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/myg6s

Van Voorhees, B. W., Fogel, J., Pomper, B. E., Marko, M., Reid, N., Watson, N., Larson, J., Bradford, N., Fagan, B., Zuckerman, S., Wiedmann, P., & Domanico, R. (2009). Adolescent dose and ratings of an internet-based depression prevention program: A randomized trial of primary care physician brief advice versus a motivational interview. *Journal of Cognitive and Behavioral Psychotherapies*, *9*(1), 1–19.

Vasilaki, E. I., Hosier, S. G., & Cox, W. M. (2006). The efficacy of motivational interviewing as a brief intervention for excessive drinking: A meta-analytic review. *Alcohol and Alcoholism*, *41*(3), 328–335. https://doi.org/10.1093/alcalc/agl016

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

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. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift Für Psychologie / Journal of Psychology*, *215*(2), 104–121. https://doi.org/ 10.1027/0044-3409.215.2.104

Xia, Y., Weng, S., Zhang, C., & Li, S. (2005). Mixture random effect model based meta-analysis for medical data mining. In P. Perner & A. Imiya (Eds.), *Machine Learning and Data Mining in Pattern Recognition* (Vol. 3587, pp. 630–640). Springer Berlin Heidelberg. https://doi.org/10.1007/11510888_62