

UCLA

UCLA Electronic Theses and Dissertations

Title

Meta-modeling social programs: Methodological reflections on a practical application

Permalink

<https://escholarship.org/uc/item/2nc16490>

Author

Lemire, Sebastian Thomas

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Meta-modeling Social Programs:
Methodological Reflections on a Practical Application

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy
in Education

by

Sebastian Thomas Lemire

2017

© Copyright by
Sebastian Thomas Lemire
2017

ABSTRACT OF THE DISSERTATION

Meta-modeling Social Programs: Methodological Reflections on a Practical Application

by

Sebastian Thomas Lemire

Doctor of Philosophy in Education

University of California, Los Angeles, 2017

Professor Christina A. Christie, Chair

Research syntheses have become an increasingly popular approach for summarizing primary research. The broad range of publications, conference presentations, journals, and books dedicated to research syntheses speaks to this point. Over the past twenty years, interest in mixed methods syntheses has steadily grown. Yet, despite the growing array of approaches for mixing methods, published applications of mixed methods syntheses are to this date few and far between. The proposed dissertation steps into this gap by illustrating the practical application of meta-modeling—a mixed methods synthesis approach. The proposed methodology structures the synthesis around the development of a meta-model and integrates findings from both effectiveness and implementation studies to determine and explain program effects. This meta-modeling does double-duty: It allows for a more nuanced appreciation of program effectiveness and allows for a better understanding of program components generating the effects.

The dissertation of Sebastian Thomas Lemire is approved.

Marvin C. Alkin

Michael Seltzer

Todd M. Franke

Christina A. Christie, Committee Chair

University of California, Los Angeles

2017

Til min Benjamin

TABLE OF CONTENTS

LIST OF FIGURES	vii
LIST OF TABLES	vii
ACKNOWLEDGEMENTS	viii
CURRICULUM VITAE	x
CHAPTER 1. INTRODUCTION	1
The Idea of Synthesizing Evidence	2
The Purpose of the Present Dissertation	4
What Follows	6
CHAPTER 2. THE LITERATURE ON MIXED METHODS SYNTHESSES	9
Summarizing the Evidence Base—Traditional Research Syntheses	9
Broadening the Evidence Base—Mixing Methods in Syntheses	15
Situating Meta-Modeling Among Mixed Methods Syntheses	18
CHAPTER 3. THE META-MODELING APPROACH	22
The Methodological Principles of Meta-Modeling	22
The Six Steps of Meta-Modeling	26
CHAPTER 4. THE CASE: ASSERTIVE COMMUNITY TREATMENT	43
The Assertive Community Treatment Model	43
The Evidence Base for Assertive Community Treatment	44
The Case for Meta-Modeling Assertive Community Treatment	50
CHAPTER 5. FRAMING THE META-MODELING SYNTHESIS	53
Step 1. Define the Scope (Research Question)	53
Step 2. Search and Retrieve Relevant Studies	56
Step 3. Conduct a Relevance Appraisal of the Studies	58
CHAPTER 6. THE IMPLEMENTATION META-SUMMARY	62
Step 4: Summarize the Critical Components in the Implementation Studies	62
CHAPTER 7. THE EFFECTIVENESS META-SUMMARY	71
Step 5: Summarize the Effect Sizes in the Effectiveness Studies	71
CHAPTER 8. THE INTEGRATED META-MODEL	88
Step 6: Develop an Integrated Meta-Model	88
CHAPTER 9. META-MODELING SOCIAL PROGRAMS—METHODOLOGICAL REFLECTIONS AND IMPLICATIONS	100
The Methodological Promises and Pitfalls of Meta-Modeling	101
The Implications of Meta-Modeling for Knowledge Building	104

Moving Forward	114
REFERENCES.....	119

LIST OF FIGURES

Figure 3.1 The structure of a meta-model synthesis.....	23
Figure 5.1 Flow chart of manuscript search and relevance appraisal.....	60
Figure 7.1 Forest plot of individual plus combined effect size (Hedges g).....	79
Figure 7.2 Forest of effect size estimates (by subgroup).....	82
Figure 7.3 Funnel plot (with trim-and-fill).....	85
Figure 9.1 Random-effects model.....	110

LIST OF TABLES

Table 2.1 Approaches to mixed-method research syntheses.....	21
Table 3.1 The six steps of meta-modeling.....	27
Table 4.1 Published reviews on Assertive Community Treatment.....	48
Table 6.1 Published fidelity measures of Assertive Community Treatment.....	64
Table 6.2 Content summary table for ACT components (abridged).....	66
Table 6.3 Five core components of Assertive Community Treatment.....	68
Table 7.1 Study characteristics (by study design).....	72
Table 7.2 Individual effect sizes (by program type).....	76
Table 7.3 Random-effects model (all 15 ACT studies).....	78
Table 7.4 Summary effect and heterogeneity (by study design).....	82
Table 7.5 Random-effects model (seven experimental studies on ACT).....	83
Table 8.1 Component coding scheme for ACT programs.....	90
Table 8.2 Component/outcome matrix for ACT programs.....	91
Table 8.3 Truth table for ACT programs.....	92
Table 8.4 Causal recipes for ACT programs (housing tenure).....	95

ACKNOWLEDGEMENTS

The list of people who have influenced and shaped my thinking on social research methods in general and evaluation more specifically continues to grow, and it is long.

For the past five years, I have benefited immensely from all the interactions and exchanges with the faculty, students, and staff that collectively comprise the Social Research Methods Division at the University of California, Los Angeles.

Above all, I wish to recognize Tina Christie who made it possible for me to pursue my PhD at UCLA (despite my “Swedish” background), generously guided me through every step of my graduate work, always found ways to challenge me intellectually, and offered me countless opportunities to grow. I came to UCLA with exceedingly high expectations, and you somehow exceeded them all. Thanks for all you do.

It goes without saying that I am also deeply indebted to the members of my dissertation committee: Marv Alkin, Mike Seltzer, and Todd Franke. Though serving on a dissertation committee is the academic equivalent of mandatory jury duty, you always went out of your way to answer questions, provide feedback, and offer guidance through every step of the dissertation. I have with the three of you on my committee been in the best of hands and minds. And I am thankful for all your time and effort.

Special thanks also go to the late Will Shadish who introduced me to meta-analysis and whose methodological guidance continues to inspire and guide me in my work. His words of statistical wisdom still ring in my ears.

I would also like to express my gratitude to Steffen Bohni Nielsen, former colleague, now friend, who knew I needed to pursue a PhD before I did, who thankfully told me, who never failed to assist me along the way, and who, like a true thought companion, continues to push my

thinking on evaluation in new directions. I hope we will keep thinking about thinking for years to come.

Over the past five years, Glory Tobiason and Nicole Mancevice, my fellow doctoral students, have become close intellectual allies and even closer friends, from whom I have learned more than words on paper could possibly express. I am beyond thankful to have shared the graduate years with both of you.

I am also pleased to thank Teresa McCarty and Mike Rose for their invaluable support over the past years. While you have not been directly involved with the present dissertation, you both represent for me the ideal combination of intellectual curiosity and integrity. Thank you for always being true role models in both words and action.

Finally, the past five years of graduate work would not have been possible without the loving and unconditional support and encouragement of Monica Alvarado Shaw. We have come a long way, and we have so many, many more exciting adventures ahead of us. I am not sure you will ever understand how deeply I love you. Thank you.

CURRICULUM VITAE

EDUCATION

M.A. Social Research Methodology, University of California, Los Angeles.	2013
M.Sc. Social Research Methods, London School of Economics	2004
B.A. Anthropology & Philosophy, University of Washington	2003

SELECTED PUBLICATIONS

- McCarty, T. L., Mancevice, N., Lemire, S., & O'Neill, H. F. (2017). Education research for a new century: a renewed vision of interdisciplinarity. *American Educational Research Journal—Centennial Issue*, 54(1), 5-22.
- Lemire, S., Christie, C. A., & Inkelas, M. (2017). The methods and tools of improvement science. *New Directions for Evaluation*, 153, 23-33.
- Inkelas, M., Christie, C. A., & Lemire, S. (2017). Value and opportunity for improvement science in evaluation. *New Directions for Evaluation*, 153, 93-102
- Christie, C. A., Lemire, S., & Inkelas, M. (2017). Understanding the similarities and distinctions between improvement science and evaluation. *New Directions for Evaluation*, 153, 11-21.
- Lemire, S., Bohni Nielsen, S., & Dybdal, L. (2012), Making contribution analysis work: a practical framework for handling influencing factors and alternative explanations. *Evaluation*, 18(3), 294-309.
- Bohni Nielsen, S., Lemire, S., & Skov, M. (2011). Evaluation capacity building: the implications of a Danish study. *American Journal of Evaluation*, 32(3), 324-344.
- Dybdal, L., Bohni Nielsen, S., & Lemire, S. (2010). Contribution analysis applied. Reflections

on scope and method. *Canadian Journal of Program Evaluation*, 25(2), 29-57.

GUEST EDITORIALS

S. B. Nielsen, S. Lemire, & C. A. Christie (2018). The Evaluation Market Place: Exploring the Evaluation Industry. *New Directions for Evaluation*.

C. A. Christie, M. Inkelas, & S. Lemire (2017). Improvement Science in Evaluation: Methods and Uses. *New Directions for Evaluation*, 153.

ACADEMIC WORK EXPERIENCE

Managing Editor, American Educational Research Journal (AERJ-SIA)	2013–16
Teaching Assistant – Causal Inference (ED255), UCLA	2015
Teaching Assistant – Research Design and Statistics Series (ED230A-C), UCLA	2014-16

SELECTED HONORS AND AWARDS

The 2016 Horowitz Foundation Social Policy Award	2016
Dissertation Year Fellowship, UCLA	2016
Research Fellow, Graduate Research Mentorship Program, UCLA	2013
The William and Louise Lucio Fellowship, UCLA	2012
The David Vickter Foundation Fellowship, UCLA	2012

CHAPTER 1. INTRODUCTION

The present dissertation rests on three premises. Premise one is that programs, projects, and policies are increasingly implemented and evaluated across a broad range of settings, populations, and times. Simply consider the many social programs that are implemented and then evaluated, only to be re-implemented and re-evaluated again. This seemingly endless evaluation cycle keeps spinning and spinning; around and around it goes, and around and around again. For the seasoned evaluator, the image of a hamster fervently spinning in its wheel may come to mind.

The second premise of the dissertation is that the vast body of evidence and knowledge produced under premise one tends to be left scattered on bookshelves like a long list of lonely and often overlooked evaluation reports. The list of ‘left-behinds’ continues to grow, and it is long. To be sure, the last evaluation report to be shelved without informing future program decisions has yet to be produced. What will continue to be produced are new additions to the massive pile of shelved, lonely, unused reports and documents. The hamster might be dead, but the wheel keeps spinning.

The third premise of the dissertation is that it is entirely possible to break this unfortunate cycle, to get the hamster out of the wheel, to make better use of the lost reports on those lonely shelves. In fact, doing so simply requires our commitment to enlist and make better use of the body of knowledge produced and contained in the still growing mountain of evaluation reports. As Glass, McGaw, and Smith (1981) pointedly remind us:

The house of social science research is sadly dilapidated. It is strewn among the scree of a hundred journals and lies about in the unsightly rubble of a million dissertations. Even

if it cannot be built into a science, the rubble ought to be sifted and culled for whatever consistency there is in it (p.11).

I agree. And in a modest attempt to avoid adding yet another pebble to the “unsightly rubble” of dissertations, the aim of the present project is to develop and extend an operational approach—labeled meta-modeling—for summarizing a broader range of evaluation ‘rubble’.

The Idea of Synthesizing Evidence

The core idea of synthesizing existing studies is well established in the social sciences. The most prevalent approach is likely the generic literature review, often provided as a setting-the-scene segment in evaluation reports and scholarly articles (Cooper, 2010; Hunt, 1997). These literature reviews, while quick and efficient to produce, often lack procedural transparency and suffer from being unsystematic and in their selection and treatment of studies, leaving them vulnerable to concerns of bias in the conclusions drawn (Cooper, 2010; Littell, Corcoran, & Pillai, 2008). As such, and even if these literature reviews may hold many merits, they are often not sufficiently rigorous to be considered research in their own right.

Another common approach for synthesizing research is that of systematic reviews. In marked contrast to traditional literature reviews, systematic reviews have the ability to systematically and transparently bring to order large bodies of research studies (Glass et al., 1981). By systematically combining the findings from a comprehensive pool of studies, systematic reviews provide robust information on program effectiveness across different settings, populations, and times. The potential value of this type of information has not been lost on policy makers, program funders, and practitioners, among others, whose interests often center on the deceptively simple question: ‘what works’? Despite the ability of systematic reviews to address

this question, one oft-cited concern relates to the exclusive (and arguably too narrow) reliance on quantitative findings from experimental (and to some extent quasi-experimental) designs—the contested winners in the “hierarchy of evidence” (Bronson & Davies, 2012).

More recently, and in a direct response to the latter concern, interest has increasingly been awarded a third approach: the mixed methods syntheses. Emphasizing the need to move beyond identifying ‘what works’, one heralded strength of these new approaches is that synthesizing a broader range of evidence allows for answers to a broader range of questions, including ‘how’ and ‘why’ interventions work (Oliver et al., 2005). More than that, and in marked contrast with traditional systematic reviews, mixed methods approaches purposely integrate qualitative, non-comparison group studies in the synthesis in their attempt to move beyond “a thin description of the evidence to produce higher order syntheses resulting in the production of new knowledge and/or theory” (Pope, Mays, & Popay, 2007). Unfortunately, and despite the many promising nature of these approaches, published applications continue to reveal a persistent and pesky challenge. As noted by Thomas et al. (2004), “integrating different types of study and data within the same review is one of the key challenges facing systematic reviewers today” (p.1012).

The persistency of this challenge, how to integrate different types of studies, cuts to the core of synthesizing existing studies; the variation in methods, designs, and findings is exactly what motivates, what prompts the need for synthesis. After all, why synthesize studies that are identical? However, this fragmented nature of existing studies also poses a formidable challenge. As formulated by Sheldon (1998):

Research studies use a variety of methods, are of variable quality and may appear to have contradictory findings ... research information can seem to users like small jigsaw pieces

in a box where there may be several pictures, several duplicates and several missing pieces (Cited in Petticrew & Roberts, 2006, p.xiii).

Sheldon's observation, while pertaining to research syntheses more generally, is only made more relevant by the recent surge of interest in mixed methods syntheses. We are—when mixing methods—not just mixing different pieces to a puzzle, but assembling, or perhaps reassembling, entirely different puzzles.

The Purpose of the Present Dissertation

Motivated by this methodological puzzle, and with the above observations as a backdrop, the following two questions guide the present dissertation:

1. How do we systematically and transparently integrate findings from different types of studies?
2. And what are the broader methodological implications of this type of integration for how we build knowledge across existing studies?

Both of these are questions about which there is considerable debate. These are also questions that are central to the credibility and promise of research syntheses. These are questions that matter for research synthesis in both theory and practice.

Towards addressing these questions, the dissertation presents and promotes the application of meta-modeling—an operational approach for mixed methods syntheses. Meta-modeling structures the integration of implementation and effectiveness studies around the development of a “meta model”—a visualization of the major trends and patterns in findings from the primary studies. Meta-modeling also relies on transparent and systematic procedures for integrating mixed evidence when developing and testing hypotheses about to what extent and in

what way programs work (or fail to work). In this way, meta-modeling speaks directly to the questions of how and in what way to integrate findings from different types of studies.

Towards advancing our practice, the dissertation will illustrate the steps in the meta-modeling approach by way of an application on Assertive Community Treatment—a care management approach for people with severe mental illness (Stein & Test, 1980). Assertive Community Treatment (ACT) is arguably one of the most well-defined and researched care management models (Rosen, Mueser, & Teesson, 2007). Many scholars and practitioners in psychiatric care consider ACT ‘evidence-based practice’ (Wright-Berryman, McGuire, & Salyers, 2011). The evidence base in support of this claim consists of a broad range of experimental, quasi-experimental, and non-experimental studies. In addition to these individual studies, a baker’s dozen of reviews have summarized the evidence base for ACT (see Chapter 4 for a more detailed examination of these reviews). Scanning across the existing systematic reviews of ACT reveals persistent support for the general effectiveness of ACT (as compared with alternative treatments) on a number of outcomes, significant variations in the effectiveness of ACT across primary studies (in ways that are not explained by current reviews), and a noticeable lack of comprehensive and systematic analyses of the “critical ingredients” in successful ACT programs (despite repeated calls for this type of information). The latter is, at least in part, due to the all-too-common focus on effectiveness studies (primarily in the form of randomized controlled trials) at the exclusion of other types of studies.

Taken collectively, then, the time is ripe for a review of ACT that (1) includes a broader range of evidence; (2) examines the unexplained variation across primary studies; and (3) provides a better understanding of the ACT program components that promote positive outcomes. The presented application of meta-modeling speaks directly to these aims.

What Follows

Grounded on the present introduction (Chapter 1), the remainder of the dissertation is structured according to nine chapters. Chapter 2 situates meta-modeling in the literature on research syntheses in general and the literature on mixed methods syntheses more specifically. While literature on research syntheses as a type of research in its own right motivates the dissertation, the proposed meta-modeling approach emerges more specifically from the literature on mixed methods syntheses. Prominent in the latter are the EPPI-Centre and realist synthesis approaches. Identifying salient similarities and differences between these approaches, the chapter concludes on the distinctive features of meta-modeling, emphasizing its intended methodological merits.

Chapter 3 considers the five methodological principles guiding the meta-modeling approach: method neutrality, division of evidentiary labor, procedural transparency, configurational causality, and commitment to learning from variation. These principles are followed by an outline of the operational procedures, that is, the six operational steps, comprising the practical application of meta-modeling. Special attention is awarded the inclusion of implementation studies, the identification of critical components on the basis of these, as well as the use of qualitative comparative techniques in developing a meta-model, as these steps represent innovative approaches in the context of research synthesis.

In Chapter 4, a brief description of Assertive Community Treatment (ACT) is provided. Building on the existing literature on ACT, the chapter first outlines the core principles and components of the ACT model, considers the major generations of research on ACT, and then awards special attention to 14 published reviews summarizing (in varied ways) the evidence base for ACT. Informed by the latter, the case is made for a meta-modeling review focusing on the critical components of ACT programs for homeless individuals.

A fruitful starting point for any methodological approach is to illustrate its real-world application. Towards this end, Chapter 5 outlines the first three steps of the approach: (1) Defining the scope and research questions, (2) searching and retrieving relevant studies, and (3) conducting a relevance appraisal of the studies. Taken collectively, these steps frame the purpose and scope of the meta-modeling synthesis. For each of these steps, all of which are applied to the ACT model, key lessons learned are provided.

Advancing the practical application of meta-modeling, Chapters 6 and 7 illustrate the separate synthesis strands of the implementation and effectiveness studies. More specifically, Chapter 6 illustrates the two cycles of coding of the implementation studies: conceptual coding (within studies) and meta-coding (across studies). The practical experiences of using these coding and analytical strategies for developing hypotheses about the critical components of ACT are considered. Chapter 7, focusing on the effectiveness studies, illustrates the effect size calculations (within studies) and the meta-analytic summary of these (across studies). The ability of this type of coding and analysis to determine whether ACT works (or fails to work) is considered.

In Chapter 8, the final integrated synthesis is presented. Grounded on Qualitative Comparative Analysis (QCA), the findings from the implementation and effectiveness synthesis strands are analytically integrated. The resulting ‘causal recipes,’ that is, the specific configurations of critical components driving the effectiveness of ACT programs, are presented and implications of these for future design and implementation of ACT programs are considered.

Chapter 9 concludes the dissertation with a set of methodological reflections on meta-modeling. First, the chapter centers on the methodological promises and pitfalls of meta-modeling, emphasizing the ability of the approach to integrate findings from implementation and

effectiveness studies in a systematic and transparent manner. Informed by the discussion of these, suggestions for future applications are made. Second, the broader implications of meta-modeling for how we build accumulative knowledge across existing studies are considered, including a call for increased emphasis on theory-driven syntheses.

CHAPTER 2. THE LITERATURE ON MIXED METHODS SYNTHESSES

Two closely related bodies of literature inform the proposed dissertation: literature on the development of research synthesis in general and literature on mixed methods syntheses more specifically. In reviewing these bodies of literature, special attention is awarded the persistent challenge revolving around how to fully integrate diverse sets of studies—a central motivation for the development of the meta-modeling approach.

Summarizing the Evidence Base—Traditional Research Syntheses

Summarizing existing research has a long and rich tradition in the social sciences (Chalmers, Hedges, Cooper, 2002). One early and oft-cited example is that of Karl Pearson's (1904) statistical summary of inoculation studies. Another noteworthy example is that of Herbert Nichols' review of 22 experimental studies of Weber's law, dating back to 1891 (cited in Petticrew & Roberts, 2006). Despite the many merits of these early contributions, research synthesis in its modern form did not gain momentum for another seven decades (Chalmers et al., 2002). Scanning across the methodological literature on traditional research syntheses, and admittedly painting in very broad strokes, reveals three discernable generations of research synthesis: the literature review, the vote counting review, and the systematic review. The following considers these in turn.

The first generation syntheses were more often than not non-statistical, relying predominantly on descriptive, narrative summaries of key findings from selected studies. These syntheses are often referred to as “qualitative,” “narrative,” or even “verbal” due to their non-systematic, selective integration of findings (Hunter, Schmidt, & Jackson, 1982). Moreover, the synthesis component often consists of a mere listing of selected findings without any—or at least

very limited—analytical integration (Hunt, 1997). The most well-established and common approach among these first generation syntheses is arguably the generic literature review, often provided as a setting-the-scene segment in research reports and journal manuscripts (Cooper, 2010; Hunt, 1997).

A number of methodological issues have over the years been raised in regards to the selection and treatment of studies in these literature reviews. First and foremost, issues pertaining to the selective and subjective inclusion of studies have been noted. Rather than aiming for comprehensive coverage, emphasis in these literature reviews is more often on ‘the best evidence’ (however defined), usually resulting in rather uncritical reviews of studies most commonly cited, closest to the researcher’s own persuasions, or simply most accessible (Shadish, 1995). An example provided by Ann Oakley (2002) is illustrative. In reviewing two existing reviews of anti-smoking education, Oakley found that the two reviews included a total of 27 studies with only 3 studies included in both reviews (2002). In addition, Oakley identified more than 70 additional studies that could have been included in the literature reviews (2002). The inability to provide any rationale as to why these reviews differ in their coverage, except for frail references to the authors’ subjective selection, is methodologically unsatisfying.

To make matters more complicated, many of these literature reviews are furthermore characterized in procedures, form, and content by personal judgment and individual preferences in the treatment of the studies included, thereby introducing another level of systematic biases (see Jackson, 1978; Pillemer, 1984). Chalmers and Lau (1994) outline the problem in this way:

Too often, authors of traditional review articles decide what they would like to establish as the truth either before starting the review process or after reading a few persuasive articles. Then they proceed to defend their conclusions by citing all the evidence they can

find. The opportunity for a biased presentation is enormous, and its readers are vulnerable because they have no opportunity to examine the possibilities of biases in the review
(Cited in Hunt, 1997, p.7)

This latter issue of limited procedural transparency, rooted in the lack of commitment to systematic methods for coding, analyzing, and integrating the studies examined, is problematic because it creates an analytical ‘black box’, leaving the reader unable to ascertain how the conclusions were drawn. To intermittently conclude, and despite their widespread and continued proliferation in evaluation reports and published journal articles, the lack of systematic and transparent procedures for the inclusion and treatment of studies, has over time served to impede the empirical credibility of these types of literature reviews.

The second generation of reviews speaks directly to these aforementioned issues. As noted by Feldman (1971), “systematically reviewing and integrating ... the literature of a field may be considered a type of research in its own right-one using a characteristic set of research techniques and methods” (p.86, cited in Cooper, Hedges, & Valentine 2009). Towards this end, the second-generation reviews pushed forward with an increased emphasis on comprehensive, quantitative summaries. Developing these summaries involved several ‘counting’ techniques:

- The vote counting method, in which studies with significantly positive, significantly negative, and non-significant outcomes are tallied. The outcome with most votes is considered the conclusive verdict (Light & Pillemer, 1984; Light & Smith, 1971).
- The accumulation of p-values in which each significance test is averaged into a pooled p-value (Rosenthal, 1978).

The main virtue of these summaries resides with their simplicity. Perhaps needless to say, both strategies also have their flaws. For instance, the lack of appreciation for different sized studies is

a serious concern related to the vote counting method. More than that, and as noted by Olkin, the “statistical Star Wars” underlying the accumulation of p-values could also be criticized on several grounds (1990). For one thing, p-values provide no information on the magnitude or direction of the effects. Leaving aside these issues, which in the mind of the present author are by no means minor, the most significant contribution of these early attempts at systematic reviews likely emerges from the pronounced emphasis on the need for systematic, transparent, and comprehensive procedures for developing reviews.

The transition from these early statistically oriented summaries to the now common meta-analytic summaries is not clearly discernable (Glass et al., 1981). However, it is safe to say that meta-analysis—the ‘analysis of analyses’—gained unprecedented momentum with Gene Glass’ Presidential Address at the Annual Meeting of the American Educational Research Association in 1976. Meta-analysis, as coined by Gene Glass (1976), was presented as the statistical pooling of findings from a comprehensive set of primary studies. The pivotal address, subsequently published in *The Educational Researcher*, was followed by the publication of Smith and Glass’ seminal meta-analysis on psychotherapy outcome studies (1977), establishing meta-analysis as the sine-qua-non of research synthesis. Slowly but steadily interest in meta-analysis grew, picking up momentum, eventually spreading across a broadening range of fields like a methodological wildfire (Cooper & Hedges, 2009; Hunt, 1997).

The popularity of meta-analyses also reached beyond academic circles to include and gain traction with policy makers and program funders. In the fervent pursuit of determining ‘what works’, the promise of systematic reviews (and meta-analytic summaries) to pinpoint effective programs both spurred and gained momentum with the ‘evidence movement’ in the early 1990s (Dixon-Woods et al., 2006). Over the past decades, the interest in systematic reviews

as part of the what works movement has not just been made visible in the rapid and enormous growth of published meta-analyses (Cooper & Hedges, 2009) but also in the establishment of (and institutionalization of systematic reviews in) the Campbell Collaboration, the Cochrane Collaboration, and later The What Works Clearinghouse—the coveted repositories of knowledge for policy and program planning. The systematic review has become part and parcel of the public sector.

There are several compelling reasons for the strong and persistent interest in systematic reviews (and meta-analytic summaries). For one thing, the ability of these reviews to systematically and transparently synthesize findings from a broad range of primary studies offers both to enhance the statistical power and generalizability of our research. Moreover, and by systematically combining the findings from multiple studies, systematic reviews also provide robust information on what works across different settings, populations, and times. This type of information potentially serves well to inform policies and practice, program planning and design.

Yet despite these encouraging aspects of systematic reviews, several criticisms have also been raised. One oft-cited concern is that systematic reviews rely too narrowly on quantitative findings from experimental and quasi-experimental designs. This methodological restriction, which is grounded on a specific definition of rigor, curtails both the types of questions posed and potential answered provided by systematic reviews, resulting in what might be termed informational ‘rigour-mortis’. To illustrate, and as argued by Pawson (2006), systematic reviews often fall short of explaining how, for whom, and under what circumstances programs work (or fail to work)—a type of knowledge increasingly prized by policy makers and practitioners alike.

In a related methodological vein, this exclusive focus on what might be termed “descriptive causation”, as captured by meta-analytic summaries of effect sizes from randomized

controlled trials, comes at the cost of limiting the ability of reviews to develop models and theories about the program mechanisms generating the observed program outcomes (Miller & Pollock, 1994; Shadish, 1996). This limitation is in the context of an increasing push towards better understanding how programs work an unsatisfactory state of affairs, prompting repeated and increasingly louder calls for a broadening of the evidence base in systematic reviews. The present dissertation reflects and responds to these calls.

In summary, the traditional approaches to research synthesis have over the past 50 years developed into an established research tradition. The emphasis, even supremacy, awarded systematic reviews (and meta-analytic summaries) of quantitative findings from experimental (and to some extent quasi-experimental) studies has served well to establish the methodological contribution and credibility of research reviews. Advancing beyond the limitations of this tradition, and especially spurred by the limited ability of meta-analytic summaries to explain how and why programs work, repeated calls have been made for a shift towards broadening the evidence base for systematic reviews and towards better understanding the mechanisms and critical components of programs.

Motivated by these calls, the proposed dissertation seeks to both ground itself and push beyond the boundaries of the traditional systematic reviews. The grounding will be in the form of adhering to the core commitment of systematic reviews by relying on systematic and transparent procedures when searching and retrieving relevant studies and by using meta-analytic techniques when developing effect size estimates (Littell et al., 2008). The position of the author is that the traditional reviews' emphasis on systematic and transparent procedures continues to hold methodological merit. The advancement beyond the traditional reviews will be in the form of broadening the evidence base for the synthesis by including and actively integrating findings

from implementation studies. The position of the author in this regard is that the inclusion of a broader range of studies allows for syntheses to answer a broader range of questions. This is admittedly a tall methodological order. Fortunately, steps in this direction have already been taken.

Broadening the Evidence Base—Mixing Methods in Syntheses

Over the past two decades, interest in mixed methods reviews has steadily grown (Bronson & Davis, 2012). As a result, a burgeoning literature showcasing methods and techniques for doing this kind of synthesis has emerged (see Dixon-Woods, Agarwal, Jones, Young, & Sutton, 2005; Saini & Shlonsky, 2012; or Sandelowski, Voils, & Barroso, 2006, for a masterful review of these). A few illustrative examples have even been published in evaluation journals. These include Scott-Little, Hamann, and Jurs (2002), combining meta-analysis with narrative review, and van der Knaap, Leeuw, Bogaerts, and Nijssen (2008), applying Campbell Collaboration standards as part of a realist synthesis. Common for these approaches is their shared aim of integrating findings from quantitative and qualitative, experimental and non-experimental studies.

In general, this new generation of mixed methods syntheses parts ways with the more common literature review by being more systematic, for example by “following formal and transparent review processes” and by “using explicit approaches to the identification and selection of evidence” (Pope et al., 2007). More than that, these reviews are also more analytically ambitious in their attempt to move beyond “a thin description of the evidence to produce higher order syntheses resulting in the production of new knowledge and/or theory” (Ibid, 2007). In marked contrast with traditional systematic reviews, these reviews purposely

include qualitative, non-comparison group studies towards enhancing the explanatory strength of the resulting reviews, emphasizing the value of a broader range of evidence (Dixon-Woods, Fitzpatrick, & Roberts, 2001).

A comprehensive presentation of the growing range of approaches to mixed methods synthesis is beyond the scope of the present work. For the present purposes, two distinct and commonly cited approaches to mixed methods reviews are worth considering in more detail: the EPPI-Centre approach and the realist synthesis approach—both of which provide the intellectual foundation for meta-modeling.

The most well-developed and empirically tested approach for mixed methods synthesis is arguably the EPPI-review (Saini & Schlonsky, 2012). Promoted by Harden and Thomas (2005; 2010), and labeled according to their affiliation with the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre) at the University of London, the approach is structured around the parallel development of individual syntheses of qualitative and quantitative evidence, subsequently merged into a combined synthesis (EPPI-Centre, 2006). The latter dovetailing of evidence represents the mixing of methods and takes the form of a thematic triangulation of data from both quantitative and qualitative studies. Following Thomas et al. (2004), this integration involves the juxtaposition of findings in a matrix, that is, the matching of “barriers, facilitators, and implied recommendations against the actual interventions that had been implemented and evaluated” (p.1011). As noted by Thomas et al., (2004) the resultant matrix allows for a better understanding of the experiences of the target groups, which in turn “could lead to the development of more appropriate and effective interventions” (p.1012).

Another prevalent mixed method approach is that of realist synthesis (Pawson, & Boaz, 2004; Pawson, Greenhalgh, Harvey, & Walshe, 2004). Developed in response to traditional

systematic reviews, the premise for Pawson's realist synthesis is the supremacy of explanatory-oriented meta-questions: Understanding how, for whom, and under what circumstances a program works (2006). Towards this aim, the realist analytical strategy revolves around the development of context, mechanism, outcome configurations (CMOs) corresponding to the underlying logic of the program under study. In its practical application, the realist *modus operandi* is to develop an initial CMO on the basis of a subset of findings, qualitative as well as quantitative, and then through iterative rounds of inclusion and synthesis of additional findings, again qualitative as well as quantitative, to refine the initial CMO configurations of the program. The underlying idea is that this step-wise, reiterative synthesis of findings will serve to refute or confirm salient aspects of the CMO, resulting in an increasingly refined understanding of how the program works.

No methodological approach is without shortcomings. Neither the realist nor the EPPI-Centre approach to mixed methods syntheses are exceptions to this rule. Whereas both approaches represent fundamental and unmatched contributions to the advancement of mixed methods syntheses, they also suffer from a lack of transparency and procedural clarity in regards to how the mixed findings are synthesized (Saini & Shlonsky, 2012). As such, practical applications of realist and the EPPI-Centre syntheses often result in an analytical 'black box' as to how findings are synthesized, and in effect how conclusions are drawn. This is unfortunate, especially given the persistent methodological challenge of how to synthesize and integrate findings from qualitative and quantitative studies in the same synthesis (Dixon-Woods et al., 2001). As such, operational clarity and transparency on this important step, and as a necessary step towards crafting credible and compelling mixed-method syntheses, is still called for.

Situating Meta-Modeling Among Mixed Methods Syntheses

The meta-modeling approach promoted in this dissertation is informed by and responds to key tenets of the EPPI-review and the realist synthesis approach. Meta-modeling shares structural characteristics with the EPPI-review in that it emphasizes distinct syntheses of quantitative and qualitative evidence, before merging these into a fully integrated mixed evidence synthesis (Harden & Thomas, 2005). The meta-modeling approach also shares much with realistic synthesis in its aim of better understanding how and in what way programs work (or fail to work). Following Pawson, it adopts the idea—yet adapts the procedures for—developing an understanding of the critical components driving the program outcomes.

In contradistinction to EPPI and realist syntheses, the meta-modeling approach specifies sequential syntheses of quantitative and qualitative studies, further specifying the latter to focus solely on implementation studies (or even fidelity measure studies). Moreover, the meta-modeling approach relies on a more structured and transparent analytical process, in which critical program components are first located within implementation studies, then summarized across these studies, and finally integrated with findings from effectiveness studies by way of Qualitative Comparative Analysis (QCA). This process is in the author's experience more operational and transparent than the procedural steps and analytical processes envisioned and prescribed by the EPPI and realist synthesis approaches.

The meta-modeling approach also departs with these approaches in other fundamental ways. One central motivation for the meta-modeling approach is to pave new way for synthesis researchers by offering more explicit, hands-on guidance on useful analytical techniques for extracting and coding causally relevant information from primary studies. The inspiration stems from qualitative coding and data display techniques developed by Miles, Huberman, and Saldaña

(2014), among others. These include specific techniques for coding, summarizing, modeling, and displaying data both within and across primary studies. The reliance on more explicit analytical techniques both enhances the explanatory strength of the approach and supports procedural transparency. The latter is also central to establishing the credibility of the conclusions drawn.

Meta-modeling also parts ways with other mixed methods synthesis approaches by focusing exclusively on implementation and effectiveness studies. This horse-for-the course approach is grounded on the idea that studies serve different purposes and in effect produce different types of information. To illustrate, experimental studies, if implemented well, provides the least biased estimates of program effectiveness. As such, awarding supremacy to these types of studies makes sense when the aim of the synthesis is to summarize program effectiveness. In contradistinction, experimental studies tend to offer limited insight into the critical ingredients of programs that generate the observed effects. For the purpose of eliciting these, for better understanding how programs work, implementation studies are more relevant. Accordingly, these types of studies are preferable when the aim of the synthesis is to identify the critical components that (individually or collectively) serve to promote the generated program effects.

Finally, the meta-modeling approach paves new methodological ground by offering operational guidance on the practical application of qualitative comparative techniques in the integration of findings from implementation and effectiveness studies. Qualitative Comparative Analysis (QCA) has received increased attention in primary research, in part because of its ability to integrate qualitative and quantitative findings in a systematic manner. However, QCA applications in the context of research syntheses are few and far between, at least in the published literature. One rare—and recent—example is that of Thomas, O’Mara-Eves, and

Brunton (2014), who applied QCA techniques on a subset of studies of community engagement in interventions for public health promotion. As noted by the authors,

There are very few examples of its use with systematic review data at present, and further methodological work is needed to establish optimal conditions for its use and to document process, practice, and reporting standards (2014, n.p.).

The present dissertation speaks directly to this call.

To intermittently summarize, the defining features of the three different mixed methods synthesis approaches are summarized in Table 2.1. The approaches are distinct in terms of seven defining features: the primary purpose of the synthesis, the scope of evidence included, the approach to study sampling, the analytical strategies for integrating findings, the type of data integration prescribed, and the key outputs of the synthesis.

Table 2.1 Approaches to mixed-method research syntheses

	EPPI-Centre	Realist Synthesis	Meta-Modeling
Purpose	Comparative: To identify patterns in data	Interpretive: To better understand how, for whom, and under what circumstances programs work.	Explanatory: To identify critical program components
Scope of evidence	Research—qualitative and quantitative.	Research and non-research sources of evidence.	Research—effectiveness and implementation studies.
Sampling	Comprehensive (Scripted)	Purposive (Non-scripted)	Comprehensive (Scripted)
Structure	Parallel.	Blended. Evidence is integrated reiteratively.	Sequential.
Analytical strategy	Meta-analytical techniques combined with thematic analysis (of qualitative data)	Reiterative refinements of program theories (unspecified)	Meta-analytical techniques combined with Qualitative Comparative Analysis
Integration of evidence	Thematic triangulation	Unspecified	Qualitative Comparative Analysis
Key output	Matrix	Program theories	Causal recipes
Examples	Harden and Thomas (2005); Oliver et al. (2005); Thomas and Harden (2008); Thomas et al. (2004),	Pawson (2006), Rycroft-Malone et al. (2012); Wong, Pawson, and Owen (2011).	Lemire (<i>under review</i>)

Table note: Adapted from Saini and Schlonsky (2012).

In summary, meta-modeling emerges from literature on research synthesis in general and literature on mixed methods syntheses more specifically. The approach is motivated by the idea of a more structured and systematic approach for summarizing findings across different types of studies. The approach is in purpose inspired by realist synthesis, in structure aligned with EPPI reviews, and inspired by Qualitative Comparative Analysis in its analytical strategy. The end product is a systematic, transparent, and operational approach for mixing methods in research syntheses.

CHAPTER 3. THE META-MODELING APPROACH

The Methodological Principles of Meta-Modeling

The meta-modeling approach is grounded on five methodological principles: method neutrality, division of evidentiary labor, procedural transparency, configurational causation, and commitment to learning from variation. There is nothing new about these principles in the broader context of the social sciences. The novelty of the principles lies perhaps more in their application in the context of systematic reviews.

A central motivation for the development of the meta-modeling approach is to broaden the evidence base for systematic reviews. Accordingly, and in line with the broader surge of interest in mixed-method syntheses, meta-modeling reflects a methodological orientation towards synthesis research, whereby qualitative and quantitative evidence are considered equally salient and valuable. Accordingly, and as shown in Figure 3.1 below, the approach includes each body of evidence, qualitative as well as quantitative, implementation as well as effectiveness data, as two sequential strands of analyses to be merged into an integrated meta-model as part of the final step of the synthesis.

In line with this broadening of the evidence base, the meta-modeling approach also relies on a firm division of evidentiary labor, a principle corresponding with a core tenet of scientific investigation: “once data have been used to develop a theory they cannot be used to test it” (Wachter & Straf, 1990, p.xxv). Following this principle, the meta-modeling approach relies on one body of evidence to generate hypotheses (findings from implementation studies) and another body of evidence to test these hypotheses (findings from effectiveness studies). By clearly demarcating the evidentiary roles, this horse-for-the-course approach optimizes the advantages

of having different types of studies, providing different types of evidence, by having them serve different—yet complimentary—purposes within the same integrated synthesis.

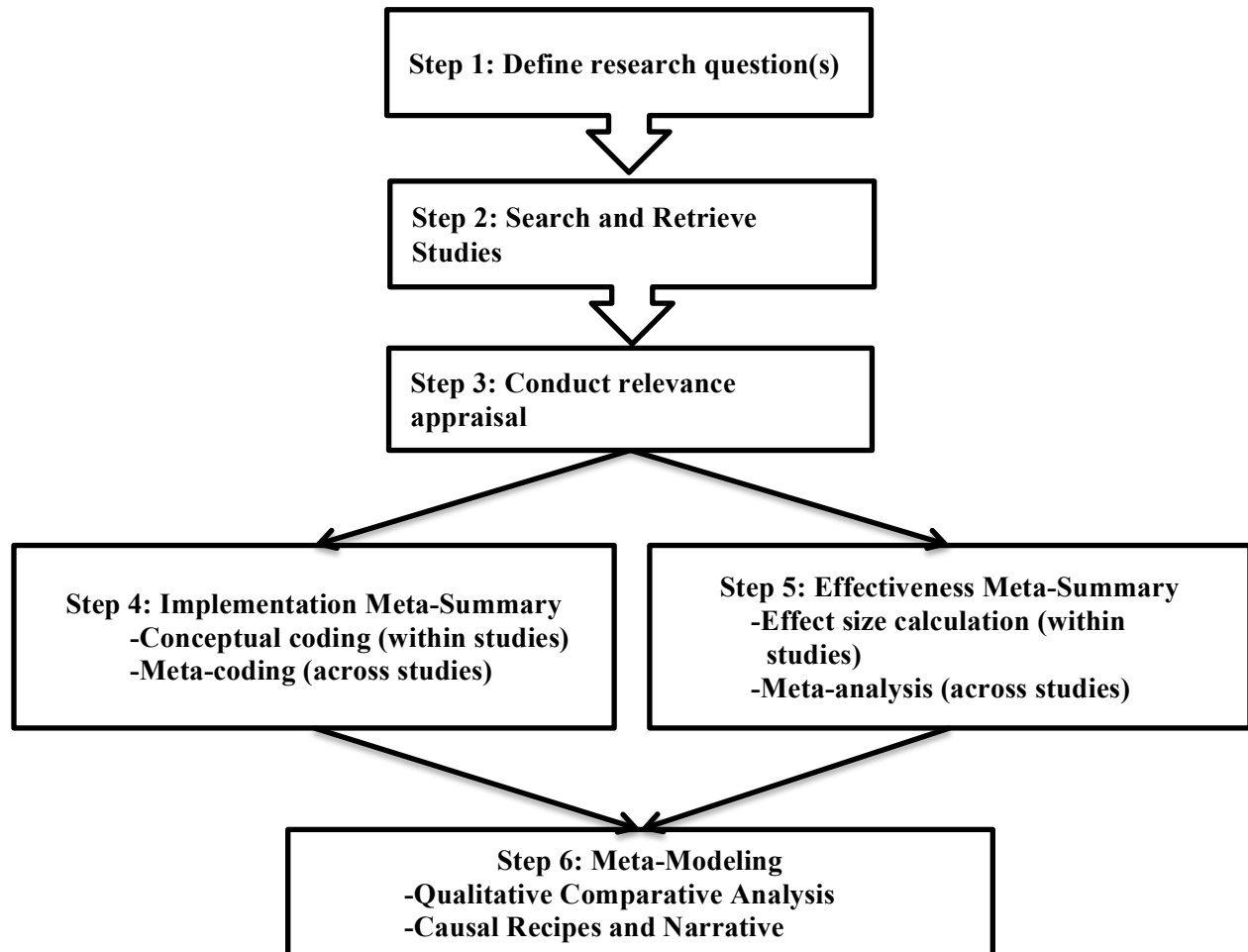


Figure 3.1. The structure of a meta-model synthesis (adapted from Harden & Thomas, 2005)

The third guiding principle cuts to the core of establishing credible conclusions on the basis of mixed methods syntheses: procedural transparency. As noted above, existing mixed methods synthesis approaches have often received criticism for lacking transparency and procedural guidance on how to integrate findings from different types of studies. The result is the

development of an analytical black box, whereby findings are integrated in ways that are indiscernible and without the promise of replication. As a result, the credibility of the conclusions drawn might be questioned. In response to this challenge, meta-modeling relies on specific coding techniques, data display methods, and analytical strategies that collectively provide an explicit chain of evidence, allowing for others to trace how studies were searched and appraised, coded and analyzed, integrated and synthesized into the final set of meta-models. This emphasis on transparency is a hallmark of traditional systematic reviews. As Glass et al. (1981) advise:

It is not uniformity in research reviewing and integrating that is desirable, rather it is clarity, explicitness, and openness—those properties that are characteristic of the scientific method more generally and which impart to inquiry its “objectivity” and trustworthiness (p.20).

The author of the present dissertation agrees. And in accordance, meta-modeling is intentionally—and in marked contrast to other mixed methods approaches—structured around a set of explicit steps and procedures. The purpose of these procedures is not to invoke mechanistic adherence to a predefined synthesis ‘recipe;’ indeed, it is strongly encouraged to adapt rather than adopt the proposed meta-modeling steps to specific research contexts. That being said, procedural transparency, which is only made more compelling when adapting research methodologies, is still to be strived for. The meta-modeling synthesis should be replicable at least in theory (Glass et al., 1981).

The fourth principle pertains to the intentional focus on causal packages: the conceptualization of causality as configurations of contributory causes that collectively, and in different constellations, lead to one or more desired outcomes (Mayne, 2012). The focus on

causal packages is grounded on the observation that many, perhaps even most, social programs (including Assertive Community Treatment) are implemented in the form of different configurations of specific program components. However, and this is where matters get complicated for systematic reviews, the specific configurations of these components often vary across different implementation settings, times, and contexts. As just one example, and as I will go on to illustrate below, ACT programs typically consist of varied constellations of a specific set of core program components, including assertive outreach, low patient-staff ratios, and immediate service provision, among other things. Without fail, the specific emphasis and constellation of these varies from program to program, from study to study. Accordingly, the treatment of ACT programs as stable and coherent wholes to which we can affix blanket causal statements (i.e., “ACT works”) simply represents a programmatic illusion that holds little relevance in the real world. Informed by this observation, one central aim of meta-modeling is to identify these critical components, the constellations of which are labeled causal recipes, driving the program outcomes.

In logical extension of the notion of causal packages, the fifth and final principle guiding the meta-modeling approach pertains to the emphasis on learning from variation. Traditional statistical techniques, including those deployed as part of meta-analyses, typically aim to control for, hold constant, or even weed out, unwanted variation. This makes sense given the analytical aim of determining the least biased net-effect estimate. In the context of systematic reviews, the approach, however intuitively appealing, often proves less than ideal. This is, at least in part, because the type of error pertaining to systematic differences across studies is exactly what motivates systematic reviews and in effect what we would want to learn from as part of our reviews. After all, if all our studies were exactly alike (in both design, implementation, and

outcomes), there would be little, if any, reason to pursue a synthesis. However, given the unavoidable diversity across programs, we would surely want to learn from, as oppose to control for, any relationship between programmatic variations and outcome patterns. For this reason, we should rather examine and learn from the variations, as opposed to controlling for it. Towards this end, the meta-modeling approach aims to learn from variation by applying Qualitative Comparative Analyses (QCA). There are several strengths to this approach. For one thing, this analytical strategy demands much fewer studies, as compared with statistical analyses. Secondly, QCA is grounded on the notion of causal packages, aiming for a better understanding of the different configurations of program components that collectively elicit positive program outcomes. This type of information is fundamentally different than the information achieved in statistical alternatives, such as meta-regression, where the primary output pertains to the relative contribution of individual program features (i.e., variables), holding other program features constant. From this perspective, meta-modeling represents an alternative approach to learning from variation.

The Six Steps of Meta-Modeling

In its practical application, the meta-modeling approach consists of six steps (outlined in Table 3.1 below). In the following, each of these steps of meta-modeling is briefly described. The first three steps—defining the scope, searching and retrieving for studies, and appraising these for relevance—correspond to commonplace steps in traditional systematic reviews (Cooper, 2010). As such, meta-modeling offers nothing new on these. Because the novel aspects of meta-modeling primarily pertain to the conceptual and meta-coding of implementation studies as well

as the use of qualitative comparative techniques in developing a final meta-model, steps four through six are covered in more detail.

Table 3.1 The six steps of meta-modeling

Step 1: Define the scope of the synthesis

- Define research question in terms of Population, Intervention, Context and Outcome (PICO standard)
- Conduct scoping review to determine scope.

Step 2: Search and retrieve relevant studies

- Define search terms and inclusion/exclusion criteria
- Conduct search for empirical papers by using multiple avenues
- Maintain a log for all the identified studies

Step 3: Conduct a relevance appraisal of the studies

- Appraise each study abstract for its relevance to the research question
- Appraise each study on the basis of full read

Step 4: Develop the implementation meta-summary

- Identify core program components (within study)
- Develop meta-summary of the critical program components (across studies)

Step 5: Develop the effectiveness meta-summary

- Calculate effect sizes (within study)
- Develop meta-summary of the program effectiveness (across studies)

Step 6: Develop the final integrated meta-model

- Integrate implementation and effectiveness findings using QCA
- Develop and describe the causal recipes for the program

Table note: Adapted from Greenhalgh, Robert, Macfarlane, Bate, & Kyriakidou (2004).

Step 1: Define the scope of the synthesis

Defining the scope of the synthesis is first and foremost anchored in carefully defining the research question(s) to be pursued. In line with more traditional review practice, the formulation of the research question(s) should adhere to the PICO standards (Higgins & Green, 2008),

specifying the Population (e.g., homeless individuals), the Intervention (e.g., Assertive Community Treatment), the Comparison (e.g., standard care), and the Outcome of interest (e.g., housing tenure). Defining these parameters of the research question serves to focus the scope and enhance the specificity of the review.

There is no hard and fast rule as to the required level of specification needed to support meta-modeling. However, a relatively broad synthesis may result in conclusions that potentially hide more than they reveal. To illustrate a synthesis of “case management” (these exist) may cover numerous competing case management models, running the risk of drawing conclusions that blur important program distinctions. In contradistinction, specifying the research questions too narrowly may result in too few studies and limited analytical opportunities to explore the findings from these, simply because the scope is overly restrictive. As such, the only guidance to provide is for synthesis researchers to carefully consider and balance the specificity of the research questions relative to the breadth and depth of the existing evidence base.

One useful strategy for defining the research question, without including too much or not enough of the existing evidence base, is to conduct a “scoping review” of published literature on the overarching topic of the synthesis (Bronson & Davis, 2012). By conducting a scoping review, the reviewer may gauge the relevance of the research questions (e.g., have the research questions already been adequately addressed by existing reviews?), identify relevant search terms (e.g., different labels for the program of interest), and even assess the feasibility of pursuing the research question given the breadth and depth of the existing literature. In the meta-modeling of ACT programs presented in chapter 4, the scoping review covered 14 existing reviews of ACT studies and helped define the research questions as well as search terms.

Step 2: Search and retrieve relevant studies

The second step of meta-modeling revolves around the search and retrieval of relevant primary studies, which generates the empirical grounding for the subsequent analyses. Two central components comprise this step: (1) the development of a comprehensive search strategy, and (2) the definition of search terms (Cooper, 2010; Card, 2012). The development of a comprehensive search strategy usually involves the reliance on multiple search modes: citation searches through electronic databases (e.g., Web of Science); manual citation searches in printed sources (e.g., existing reviews); consultation with subject matter experts; and/or open electronic searches (e.g., Google scholar). As Sandelowski and Barroso (2007) remind us, “the most important threat to the validity of any research synthesis—whether of qualitative or quantitative findings—is the failure to conduct a sufficiently exhaustive search” (p.35).

The specification of relevant and effective search terms for the electronic search is equally important. Informed by the scoping review, and corresponding to the formulated research questions, the search terms should carefully balance the potential for long lists of false positives (studies falsely identified as relevant in the electronic search) versus failing to identify and in effect include relevant studies. There is no assured recipe for achieving this balance. The only solution is to approach the search process as an iterative, trial-and-error process, merging back and forth between electronic searches and refinement of search terms and strategies.

Despite the iterative nature of the search process, the final search procedure is still to be systematic and documented for the purpose of transparency and possible replication (Bronson & Davis, 2012). Heeding this advice, the development of a search and retrieval log, tracking and

detailing the number of manuscripts (duplicate and unique records) identified by source (e.g., CINAHL), is in the author's experience an essential component of the search and retrieval process.

Step 3: Conduct a relevance appraisal of the studies

The third step in the meta-modeling synthesis is a relevance appraisal. Central to this step is the formulation of clear inclusion/exclusion criteria. These criteria may include program features (e.g., ACT program combined with peer advisor); outcomes of interest (e.g., housing tenure); target group (e.g., homeless individuals); study design (e.g., experimental studies); time frame (e.g., past ten years); and publication type (e.g., peer-reviewed articles, books, or unpublished dissertations). The formulation of and rationale for these types of inclusion/exclusion criteria are illustrated in the ACT example provided below. For now, it suffices to say that the definition of the inclusion/exclusion criteria in tandem with the formulation of the research questions collectively serve to demarcate (more than anything else) the boundaries of the synthesis and in effect the conclusions drawn. For that reason, the definition of these boundaries should be carefully aligned with the PICO parameters defined for the research questions and established prior to any analyses of any identified studies (in step 4 and onwards).

In marked distinction with current guidance on review practice, the meta-modeling approach does not encourage nor discourage relevance restrictions in terms of publication date, location, or context of the studies (Card, 2012). In fact, generic guidance or seemingly hard and fast rules about what always to include or exclude in any synthesis run the risk of promulgating unproductive mechanistic decision-making at the expense of methodological reflection. The latter should always hold supremacy at any stage of the synthesis.

One useful strategy for relevance appraisal is to structure it sequentially around two stages. In the first stage, the abstract for each identified manuscript (e.g., journal article, dissertation, or book) is carefully reviewed. Informed by the abstract review, manuscripts are categorized according to three categories: (1) Excluded (e.g., clearly out of scope or inapplicable format); (2) potential study for the implementation meta-summary (relevant for the implementation meta-summary); and (3) potential study for the effectiveness meta-summary (relevant for the effectiveness meta-summary). Illustrations of each of these categories are provided in the worked ACT example in Chapter 5. The overarching aim of this initial relevance appraisal is two-fold: to identify obviously irrelevant studies (i.e., false positives) and to categorize the remaining manuscripts according to their analytical purpose in the synthesis. To minimize the risk of excluding a potentially relevant study on the basis of an abstract, one should always err on the side of inclusion, even if the potential relevance of the manuscript appears faint. The worst error at this stage of the synthesis is to exclude a useful study.

Informed by the first stage of relevance appraisal, the second stage focuses on the subset of manuscripts categorized for the implementation and effectiveness meta-summaries. Based on a full read of these manuscripts, the primary purpose of this appraisal is informational relevance specific to either the implementation meta-summary or the effectiveness meta-summary. Accordingly, and given the different evidentiary roles of these meta-summaries, the specific inclusion criteria for each of these differ in important ways.

For manuscripts considered for the implementation summary, the inclusion criteria primarily serve to identify those relevant for the identification of critical program components, that is, program components that serve as drivers of the program outcomes. These studies may be in the form case studies of program implementations, model program descriptions (formulating

program principles in general terms), or ideally studies reporting in the development and empirical testing of fidelity measures. Accordingly, the inclusion criteria specify (1) the presence of a specific program type (e.g., Assertive Community Treatment) and (2) a distinct focus on program adherence, integrity, fidelity, or degree of implementation. Studies that fail on one or both of the inclusion criteria are excluded for further analysis. Further inclusion criteria, pertaining for instance to study design, subpopulation, outcomes, or context, may be considered but are not required. This is because the purpose of the implementation meta-summary is to generate generic hypotheses about the critical program components. Because the purpose of the implementation meta-summary is to develop hypotheses about the critical components, as opposed to testing these, the inclusion criteria are less restrictive than those of the effectiveness meta-summary.

For the candidate manuscripts for the effectiveness meta-summary, the inclusion criteria aim to identify manuscripts that support effect size calculation for a specific program, population, comparison, and context (as defined by the research question). Accordingly, the inclusion criteria always specify a study design with a control or comparison group (requisite for effect size calculation) and demand enough information to support effect size calculation (e.g., means and standard errors). In addition, and depending on the specific formulation of the research question (recall PICO criteria defined in the framing stage of the synthesis), the inclusion criteria may furthermore specify the presence of a specific program variant (e.g., ACT as opposed to intensive case management program), one or more relevant outcomes (e.g., housing tenure), and the inclusion of a specific target group (e.g., homeless individuals with a dual diagnosis), or a specific setting (e.g., urban environments). Because the studies included in the effectiveness summary will serve to test the critical component hypotheses, the specific

inclusion criteria should emerge from and be aligned with the research question and the PICO criteria defined in the scoping stage of the synthesis. Studies that fail on one or more of the inclusion criteria are excluded for further analysis.

The end product of the relevance appraisal, then, is in the form of two subsets of manuscripts for further analysis: one subset relevant for the implementation meta-summary and one relevant for the effectiveness meta-summary. Because the development of these distinct meta-summaries (steps 4 and 5) comprises the methodological backbone of the meta-modeling approach, a more detailed description of these steps will be provided in what follows.

Step 4: Develop the implementation meta-summary

The overarching aim of the implementation meta-summary is to generate hypotheses about the critical program components promoting one or more specific outcomes. The development of the meta-summary is structured around two cycles of coding: conceptual coding (within primary studies) and meta-coding (across primary studies).

Conceptual coding. The purpose of the first cycle of coding is to identify and describe a comprehensive set of program components on the basis of each of the included implementation studies. As noted above, the manuscripts may represent a broad range of implementation studies, including manuscripts covering the conceptual and empirical development of program fidelity measures, empirical applications of these, or case studies of program implementations, to name but a few. As I will go on to illustrate in the ACT example (Chapter 6), it might be worth awarding particular attention (or even exclusive attention) to studies developing fidelity measures, as these are often more specific and empirically grounded in their articulation of core components comprising the program under study. Moreover, the fidelity measures developed in

these manuscripts often serve as the conceptual foundation of the broader range of implementation studies.

Towards summarizing the identified program components, one useful data display technique is that of a content matrix (see Chapter 6 for an illustrative example). The matrix simply lists the identified components by each manuscript. More specifically, the matrix specifies the program component as labeled in the manuscript, and if needed, a specification of any nuances in meaning (i.e., definition) across the included manuscripts. This anchoring of each component with the language from the individual manuscript both provides analytical depth to the components and provides a transparent chain-of-evidence that allows other researchers to examine the grounding for the coding of the components. In summary, the three-fold purpose of the matrix display is to (1) provide an overview of the different components contained within each of the manuscripts, (2) facilitate identification of patterns across these components, and (3) provide transparency on the coding of the manuscripts. This latter point is important for the purpose of methodological transparency. An example of a completed content matrix is provided in Chapter 6.

Meta-coding. On the basis of the content matrix, the second coding cycle aims to identify the most salient, that is, the most critical components of the program under study. This identification demands synthesis across the program components included in the content matrix. Towards this aim, one useful strategy is to scan across the codes and identify broader categories or types of components. These component types may, for instance, be in the form of program structure, program philosophy, or program services. These categories serve well to reorganize the individual components identified within each implementation study, allowing in turn for cross-study patterns to emerge more clearly. Once the individual components have been

reorganized, the analytical strategy is to identify the most prevalent components across the included studies (Miles et al., 2014).

The final meta-summary is in the form of a schematic overview of the identified set of critical components supported by an operational definition for each and a specification of how to rate the presence of the components in the effectiveness studies (see Chapter 6 for an illustrative example). The underlying idea of this type of meta-summary is that it clarifies the meaning of each critical component and provides grounding for the subsequent coding of these critical components in the effectiveness studies.

Taken collectively, the identified and operationalized set of critical components serves to structure and inform the later development of causal recipes for the program under review. This deeper analysis will be the focal point of the final meta-modeling synthesis, integrating information from the effectiveness and implementation meta-summaries by way of qualitative comparative techniques. Before advancing this final analysis, however, another type of synthesis, focusing on program outcomes, has to be developed. This synthesis, the effectiveness meta-summary, is developed in the fifth step.

Step 5: Develop the effectiveness meta-summary

In parallel structure to the implementation meta-summary, the effectiveness meta-summary consists of two stages of coding and analysis: within and across primary studies.

First coding cycle: estimate effect sizes. The first cycle of coding of the effectiveness studies focuses on salient background information from each study, including publication information (author name, publication year); study features, , type of control (e.g., treatment as usual or a specific alternative program), and study design (e.g., experimental or quasi-

experimental); program design features (e.g., the absence or presence of specific program components); and information relevant for calculating effect sizes such as group means for treatment and control, sample sizes for these (after attrition), and corresponding standard errors.

The specific codes used, especially those pertaining to program features, are contingent on the nature of the program and the research question(s) posed. For instance, if the research question guiding the review pertains to specific program variations (e.g., assertive outreach teams) that are tailored to a specific subpopulation (e.g., homeless individuals), these features would have to be coded for further analysis.

On the basis of this first coding cycle, individual effect sizes for each study are calculated using the Practical Meta-Analysis Effect Size Calculator (Lipsey and Wilson, 2001). The effect size measure calculated is the standardized mean difference statistic (Cohen's d) corrected for small sample bias using a Hedges' g correction (Hedges & Olkin, 1985); though, alternative effect size measures work equally well (e.g., correlation coefficients, odds-ratio, or risk-ratio). The primary purpose is to obtain a standardized effect size for each salient outcome and comparison, as determined by the scope of the synthesis.

Second coding cycle: meta-analysis. The purpose of the second stage of analysis is to determine the overall effectiveness of the program under study, to gauge the degree of unexplained variation between the studies, and to assess the presence (and influence) of publication bias. Towards these aims, and on the basis of the individual effect sizes, a random-effects model is fitted to estimate a combined effect size across all comparison group studies (Lipsey & Wilson, 2001). In estimating the combined effect size, inverse variance weighting is applied to give more weight to the more precise estimates (i.e., studies with low sampling error).

One common meta-analytic data display technique is that of a forest plot, presenting the individual and combined effect size estimates along with corresponding confidence intervals. The combined effect size should be estimated separately according to experimental or quasi-experimental study designs, and perhaps even broken down by other relevant design or contextual features, depending on the relevance and distribution of these features in the sample of studies. To illustrate, the worked example on ACT studies presented below considers effect sizes for different study designs and program types, both of which are considered relevant when assessing the average impact of ACT programs.

In advancing the analyses even further, and as an integral part of examining the combined effect size, attention is then awarded the unexplained variation (also referred to as heterogeneity) between the included primary studies. Examining the amount of unexplained variation is important for several reasons. First and foremost, the presence of large amounts of unexplained variation jeopardizes the accuracy of the effect size estimates, as reflected in widened confidence intervals. Secondly, yet of no less importance, unexplained variation affects our ability to meaningfully interpret the combined effect size estimate. This is because the unexplained variation may be due to cross-study variations in program design features, contextual conditions, target group characteristics, or other pertinent particularities, that in some way affect the effectiveness of the programs being studied.

Towards examining the unexplained variation, and in line with the traditional meta-analytic framework, meta-modeling seeks to gauge both the significance and the volume of unexplained variation. Accordingly, the unexplained variation is examined using the common I^2 statistic promoted by Higgins et al., (2002). The presence of unexplained variation, as

exemplified in the worked example presented below, is not necessarily a detrimental to the review. From the perspective of meta-modeling it simply motivates further analyses.

Another central component of the effectiveness meta-summary is the assessment of publication bias. Publication bias, also referred to as the file-drawer problem, pertains to reviews that solely rely on published studies (the present review included). The potential bias emerges from non-significant findings being less likely to be published (or reported in publications), as compared with significant findings (see Easterbrook, Berlin, Gopalan, & Matthews, 1991; and Dickersin, Min, & Meinert, 1992, for examples of this). From the perspective of effect size estimation, this publication bias is problematic because the missing studies are systematically different from the included studies in a way that leads to an overestimation of the combined effect size, especially if the studies are characterized by relatively small sample sizes (Borenstein, Hedges, Higgins, & Rothstein 2009). As such, and in the presence of publication bias, the combined effect size estimates will be upward biased (overstating the program effect).

As with unexplained variation, there are many different ways to assess publication bias. For the purpose of meta-modeling, emphasis is awarded diagnostics capturing the potential presence and influence of publication bias. One useful diagnostic to assess the presence of bias is the funnel plot, depicting the individual effect size estimates (x-axis) by their standard errors (y-axis). An example based on the ACT studies is provided below. The presence of publication bias is in the funnel plot indicated by a non-symmetrical distribution of the individual effect size estimates around the mean effect size (Borenstein et al., 2009). If the pattern suggests the presence of bias, the logical next step is to consider its relative influence on the combined effect size estimate. One approach is the trim and fill procedure, which accounts for the effect of publication bias by re-estimating a balanced (i.e., unbiased) effect size estimate. If the difference

between the unbiased and biased estimate is minor, the concern for publication bias can be considered trivial.

The purpose of the effectiveness meta-summary is twofold. First and foremost, the meta-summary serves to develop individual effect size estimates, which, in combination with the critical components identified in the implementation meta-summary, will comprise the building blocks for the final meta-modeling synthesis. Secondly, the effectiveness meta-summary provides one or more combined effect size estimates that respond directly to the “what works” question of meta-modeling. Accordingly, the meta-summary both serves to answer a central question for the meta-model synthesis as well as to provide the effect size estimates for the final meta-model.

Step 6: Develop the final integrated meta-model

The final step in the meta-modeling synthesis is the integration of findings from the implementation and effectiveness meta-summaries. The modus operandi of the integrated synthesis, the way in which the implementation and effectiveness findings are integrated, is grounded on Qualitative Comparative Analysis (QCA). QCA is a comparative analytical technique that seeks to identify the sets of causal conditions (termed causal recipes) that trigger a specific outcome (Ragin, 2014; Schneider & Wagemann, 2013). In the present synthesis, QCA allows for the identification of causal recipes that promote high program performance in relation to specific outcomes. Informed by Rihoux and Ragin (2009), the QCA involves six steps:

1. Recode effectiveness studies according to the critical components identified in the implementation studies (component coding);
2. Develop a component/outcome matrix;

3. Use QCA software to create a ‘truth table’;
4. Minimize solution;
5. Resolve contradictory configurations;
6. Present final interpretation of solutions.

In the first step, each of the included effectiveness studies are recoded according to critical components identified as causally salient in the implementation meta-summary (i.e., the final set of core components). These codes represent salient components that might explain how and why the program works (or fails to work). A note on this assessment and coding is called for. In traditional crisp-set QCA, cases (i.e., the effectiveness studies in the meta-modeling approach) are coded on a binary scale, whereby a “zero” or a “one” denotes the absence or presence of a given component. Leaving the appealing simplicity of this binary coding aside, this type of coding does not reflect that the presence of a given critical component is more likely one of degrees. Accordingly, the extent to which each of the components is present in the studies is better assessed according to multiple levels, such as: 0 (no presence), .33 (low presence), .67 (high presence) and 1 (full presence). This coding is termed a fuzzy-set in QCA terminology.

Perhaps needless to say, no hard and fast rules can be made regarding the decision to use crisp or fuzzy-set coding, a decision that is entirely contingent on the nature of the program reviewed and the level of program description in the available studies, among other things.

In the second step of the QCA, the coded studies are then organized and displayed in a table format and linked with recoded outcome values (based on the estimated effect sizes). See Table 8.1 in Chapter 8 for an illustrative example. This component/outcome matrix is what serves as the primary input for the QCA package to integrate and carry out subsequent analysis.

On the basis of the component/outcome matrix, the logical configurations of critical components that elicit a positive outcome can be identified and presented in a truth table, each row representing a specific configuration of components (i.e., ‘causal recipe’) that elicits a positive outcome. For each row, that is, for each causal recipe, the proportion of studies contributing to the recipe is also presented. Several software packages for conducting QCA can be used for this purpose. In the ACT example presented below, FsQCA, a software package developed by Ragin & Davey (2014), was used.

This part of the analysis may already reveal a number of interesting insights about the inner workings of the program being studied. First and foremost, the truth table will reveal the array of causal configurations present in the dataset (i.e., the causal recipes represented in the effectiveness studies). The specific configurations revealed, as well as their coverage across the existing studies, may provide valuable insight about common inclinations in program design. In addition, logical remainders, that is, causal configurations not present in the dataset, are also made explicit.

The final analytical step, however, is where the final causal recipes are revealed. In this step, the FsQCA software applies inferential logic (Boolean set algebra) to simplify the truth table into the causal recipes that are minimally sufficient to produce a positive outcome. The results are summarized in a table format, displaying the identified causal recipes, their coverage in the existing studies, and the level of consistency of the causal recipe (i.e., the ratio of studies displaying the causal recipe accompanied by a positive outcome relative to studies displaying the causal recipe unaccompanied by a positive outcome).

In the presence of low levels of consistency for the derived causal recipes, a reiterative process involving reexamination of the studies, and perhaps even introduction of new conditions

to resolve any inconsistencies, will be applied to identify the configurations that best explain the outcome, with the least amount of inconsistency. The final “set” of causal recipes are the causal recipes that on the basis of implementation and effectiveness findings produce high performing programs.

Develop the final meta model(s). Informed by the findings from the QCA, the final step consists of presenting and describing the causal recipes. Accordingly, a summary of these causal recipes is provided in both visual and narrative form. Illustrative examples are provided in Chapter 8 below. To enrich the explanatory power of the causal recipes, information about the critical components in the implementation meta-summary may be re-examined.

CHAPTER 4. THE CASE: ASSERTIVE COMMUNITY TREATMENT

The following chapter serves to set-the-scene for the subsequent meta-modeling application on Assertive Community Treatment (ACT) programs. Towards this aim, the first section of the chapter describes the background, underlying principles, and main characteristics of the ACT model. In extension of this description, the second part of the chapter considers the substantial and still growing body of evidence on ACT programs, awarding particular attention to 14 published reviews of ACT studies. Informed by these reviews, the chapter concludes with a set of reflections on the need for a systematic review that (1) examines the variation in program effectiveness in relation to housing tenure; (2) provides a better understanding of the critical components that promote positive housing tenure; and (3) integrates a broader range of studies—implementation and effectiveness—on ACT programs for homeless individuals.

The Assertive Community Treatment Model

The Assertive Community Treatment (ACT) model, also known as Program of Assertive Community Treatment (PACT), is a prevalent care management approach for people with severe mental illness (Stein & Test, 1980). Emerging in the 1980s, ACT is arguably one of the most well-defined and well-researched care management models (Bond, Drake, Mueser, & Latimer, 2001; Rosen et al., 2007). The core principles guiding the ACT model are: (1) low patient to staff ratio; (2) services provided in the community; (3) shared caseloads across a multidisciplinary team; (4) 24-hour coverage; (5) services provided directly by the care team; and (6) no time limit on service provision (Mueser, Bond, Drake, & Resnick, 1998). These ACT-principles are described and illustrated in seminal articles by Test (1992) and Witheridge (1991), among others,

and generally recognized as central, if not defining, features of ACT programs by both practitioners and scholars (Bond, McGrew, & Fekete, 1995).

In their practical implementation, ACT programs typically target individuals experiencing severe mental illness (defined in different ways) and characterized by extended or frequent hospitalizations or use of emergency services, persistent symptoms, co-occurring substance use, criminal justice involvement, or homelessness (Salyers et al., 2010). These difficult to treat populations are offered a broad range of services, including, but not limited to, medical, psychosocial, practical, social functioning, and rehabilitative support (Dixon, 2000). A multidisciplinary team makes these services available seven days a week, 24 hours a day; though, the frequency, duration, and intensity of services may depend on client preferences (Stein & Test, 1980). The services are individualized and provided in-vivo where the client lives, works, and socializes (Marx, Test, & Stein, 1973; Test, 1992). Despite the widespread agreement on the core principles and features of the ACT model, the implementation of these are often adapted and tailored to local settings and populations—a broader point of discussion I will return to in Chapter 9.

The Evidence Base for Assertive Community Treatment

Many scholars and practitioners in psychiatric care consider ACT ‘evidence-based practice’ (Bond et al., 2001, among others). As noted by Burns and Santos (1995), “research has been a hallmark of the development of assertive community treatment programs...” (p.669). In a similar vein, Dixon (2000) declares ACT the “gold standard for creative program development, rigorous evaluation, and ongoing adaptation” (p.763). Assertive Community Treatment is not just effective practice, or better practice—but best practice (adapted from Patton, 2001).

The evidence base enlisted in support of this coveted claim consists of a broad range of experimental, quasi-experimental, and non-experimental studies. Over the past 30 years, a steady flow of studies have illustrated favorable outcomes of ACT in relation to treatment retention (Santos et al., 1993), hospital admission rates (Chandler, Meisel, McGowen, Mintz, & Madison 1996), duration of hospital stay (Quinlivan et al., 1995), and to some extent housing retention (Lehman, Dixon, Kernan, DeForge, & Postrado 1997), employment (McGrew, Bond, Dietzen, McKasson, & Miller 1995), and client costs (Jerrell & Hu, 1989), among other outcomes.

Primary studies of ACT have also covered a broad range of populations, including people with dual diagnosis (Drake et al., 1998), people with substance abuse (Teague, Drake, & Ackerson, 1995), recently incarcerated (Solomon and Draine, 1995), veterans (Rosenheck, Neale, Leaf, Milstein, & Frisman, 1995), and people experiencing homelessness (Morse, Calsyn, Allen, Tempelhoff, & Smith 1992; Morse et al., 1997). Moreover, studies have covered program implementations in urban (Dixon, Friedman, & Lehman, 1993) and rural settings (Fekete et al., 1998), and, reflecting the dissemination of ACT programs internationally, examined ACT implementations in Sydney (Hoult & Reynolds, 1984) London (Audini, Marks, Lawrence, Connolly, & Watts, 1994), and Ontario (Lafave, deSouza, & Gerber, 1996), among other places. The evidence base is broad.

On the cusp of the early waves of effectiveness studies, around the mid 1990s, and parallel to the broadening dissemination and adaptation of the ACT model, researcher and practitioner focus increasingly centered on program fidelity, resulting in a stream of published studies on ACT model adherence, critical ingredients, and fidelity of implementation (McGrew, Bond, Dietzen, & Salyers, 1994; Monroe-DeVita, Teague, & Moser, 2011). Spurred in part by concerns of program drift (local ACT adaptations veering away from the original ACT

principles) and in part by the promise of establishing ACT as evidence-based practice, the persistent focus of these studies revolves around conceptual and empirical examinations of the critical components comprising ACT programs (Monroe-DeVita et al., 2011). Citing Bickman (1987), McGrew et al. (1994) comment: “A crucial, but often ignored step in documenting program implementation is specification of the critical components for a given model” (p.670). In line with this thinking, and serving as a theoretical backbone for many of the implementation studies, five successive frameworks for measuring ACT program fidelity were published (McGrew et al., 1994; McGrew & Bond, 1995; Teague et al., 1995; Teague, Bond, & Drake, 1998; and Monroe-DeVita et al., 2011), each offering slight variations in their program composition of ACT (see Chapter 6 Developing the Implementation Meta-Summary for a more detailed discussion of these).

In addition to the individual effectiveness and implementation studies, numerous reviews have summarized the evidence base for the ACT model. In preparing the present review, a total of 14 research reviews with an exclusive focus on ACT were identified (see Table 4.1 below). Most of these reviews were published in the late 1990s (between 1995 and 2000), reflecting perhaps a coming of age moment for the ACT evidence base. The most recent reviews were published in 2007. Considered collectively, the reviews cover ACT-studies published in the time period from 1973 to 2007, with several reviews overlapping in their coverage. The number of studies included in the reviews range from 5 studies (Marshall & Creed, 2000) to 34 studies (Latimer, 1999).

Across most of the reviews, focus is exclusively awarded experimental and quasi-experimental studies, reflecting an emphasis on summarizing and establishing the effectiveness of ACT. Whereas seven reviews rely on narrative summaries to describe positive outcomes of

ACT, five reviews involve meta-analytic techniques, one of which is a Cochrane review (Marshall & Lockwood, 1998). Whereas the scope in most of these reviews considers ACT studies in general, a few studies focus more specifically on variants of ACT, including Assertive Outreach teams (Bond et al., 1995), the Program of Assertive Community Treatment (Herdelin & Scott, 1999; Olfson, 1990; Taube, Morlock, Burns, & Santos 1990), ACT-studies in the UK (Marshall & Creed, 2000), or ACT programs for homeless individuals (Coldwell & Bender, 2007).

Of particular salience to the present dissertation, six of the reviews report—with varying degrees of emphasis—on housing-related outcomes among ACT participants. Of these, one review is in the form of a meta-analysis of ten experimental and non-experimental studies on ACT programs for homeless individuals. Scanning across these reviews reveals a number of interesting trends. First and foremost, the effectiveness of ACT in enhancing housing tenure is generally positive; though, varying markedly across primary studies in ways that are not explained by current reviews (Coldwell & Bender, 2007). In a related vein, the “critical ingredients” in successful ACT programs, while repeatedly called for, have not been identified (Burns and Santos, 1995; Marshall & Lockwood, 1989; Rosen et al., 2007). Finally, while several of the existing reviews rely on a mix of experimental, quasi-experimental, and non-experimental studies, the analysis of these is treated as separate (non-integrated) strands of synthesis.

To intermittently conclude, and despite the many merits of these existing reviews, a number of shortcomings exist. For one thing, the noticeable variations in program effectiveness, compounded by the generic focus on ACT programs generally speaking, makes it difficult to reach any definitive conclusions about the effectiveness of ACT. The amount of unexplained

variation is simply too severe. To make matters more complex, the almost exclusive focus on effectiveness studies, at the expense of including implementation studies, leaves the reviews unable to examine how diversity in program designs, target groups, and settings influence (i.e., account for) the observed outcome variations.

To move the knowledge base on ACT forward, then, future reviews would have to have a stronger theory-driven orientation (shifting away from solely focusing on *whether* or not ACT works and towards theorizing about *how* ACT works), provide a sharpened focus on specific subpopulations (allowing for a more fine-grained understanding of whom ACT works for), and broaden the evidence base to include studies on critical program components driving the program outcomes (covering both effectiveness and implementation studies). Of the currently published reviews, none provide this sort of information; yet understanding how and for whom ACT works is precisely the type of information that is needed to begin to address questions concerning the effectiveness of emphasizing (or deemphasizing) specific program components when designing and implementing new ACT programs.

Table 4.1 Published reviews of Assertive Community Treatment

Author (Year)	Number of ACT studies	Review Type	Scope	Findings (housing outcomes)
Olfson (1990)	11 experimental studies (1973-89)	Narrative	PACT	Not reported
Taube et al. (1990)	5 experimental studies (1980-89)	Narrative	PACT (cost-effectiveness)	Not reported
Test (1992)	8 experimental studies (1973-92)	Narrative	ACT	Not reported

Burns & Santos (1995)	8 experimental studies (1990-94)	Narrative	ACT	ACT reduced days homeless (1 study)
Bond et al. (1995)	9 experimental, quasi-experimental, and non-experimental studies (1982-93)	Meta-analysis	Assertive Outreach (Thresholds Bridge)	Not reported
Marshall & Lockwood (1998)	20 experimental studies (1973-98)	Meta-analysis	ACT	ACT reduced proportion homeless (3 studies) and increased days in housing (1 study)
McHugo et al. (1998)	13 experimental, quasi-experimental, and non-experimental studies (1995-98)	Narrative	ACT (methodological issues)	Not reported
Latimer (1999)	34 experimental and quasi-experimental studies (1982-98)	Narrative/multiple regression	ACT	Not reported
Herdelin & Scott (1999)	19 experimental studies (1973-97)	Meta-analysis	PACT	Housing integrated in social functioning measures
Marshall & Creed (2000)	5 experimental studies (1994-98)	Meta-analysis	ACT (UK)	Not reported
Dixon (2000)	46 Experimental, quasi-	Narrative	ACT	Not reported

	experimental, and non-experimental studies (1973-99)			
Bond et al. (2001)	25 experimental studies (1980-99)	Narrative	ACT	ACT improved (8 studies), made no difference (3 studies), or worsened (1 study) housing stability.
Coldwell & Bender (2007)	10 experimental and non- experimental studies (1992-2000)	Meta-analysis	ACT (homeless persons)	ACT studies demonstrated a 37% reduction in homelessness
Rosen et al. (2007)	43 experimental, quasi- experimental, and non-experimental studies (1980- 2007)	Narrative	ACT	“roughly half” of ACT studies improved housing stability

The Case for Meta-Modeling Assertive Community Treatment

There are several reasons for meta-modeling Assertive Community Treatment (ACT) in general and for focusing on ACT programs for homeless individuals more specifically. First and foremost, and as indicated above, the current meta-analytic and narrative reviews have primarily served to summarize the effectiveness of ACT, most often by focusing exclusively on quantitative findings from experimental and quasi-experimental effectiveness studies. While these reviews have served well in this regard, none of the reviews have sought to systematically

integrate findings from a broader range of studies (e.g., implementation studies). Accordingly, a significant part of the evidence base for ACT has as of writing not been systematically reviewed.

Secondly, the findings from existing reviews raise fundamental questions about the effectiveness of ACT. While several reviews document persistent positive findings on a number of outcome domains (e.g., hospitalization and patient retention), noticeable variation in effectiveness is also persistent across studies, especially in relation to housing tenure (the outcome domain examined in the present dissertation). This variation, while documented in the three most recent reviews of ACT, has not been examined systematically in any of the present reviews. Accordingly, a better understanding of the variation across program settings and contexts is called for.

Third, and in part emerging from the documented variation in effectiveness across studies, repeated calls have been made for better understanding the critical ingredients of ACT programs. As noted by Rosen et al. (2007), among others, the relative importance of specific ACT program components is “presently unknown and a key issue for future research” (p.815). Fortunately, the ACT model is a highly formalized and researched care management model, why it lends itself particularly well to a synthesis aiming to elucidate critical program components.

Fourth, and as noted by McHugo et al. (1998), “there is no well-articulated and comprehensive theory of ACT from which to draw strong hypotheses for experimental evaluations” (p.247). Accordingly, and as McHugo et al. (1998) go on to lament, “there is a conspicuous lack of theory-driven meta-analyses of ACT studies” (p.247). The theoretical grounding, underlying the ACT model, and potentially supporting a better understanding of how it generates positive outcomes, has not been articulated or examined as part of the existing reviews. This is unfortunate, at least in part, because it keeps us from testing more specific and

arguably more informative hypotheses about the effectiveness of ACT. In agreement with McHugo et al. (1998), a more theory-oriented synthesis, such as meta-modeling, is called for.

Fifth and finally, the present focus on homeless individuals with mental illness emerges from the persistent challenge of providing care for this subpopulation. As noted by Coldwell and Bender (2007), an estimated 14 million Americans experience one or more episodes of homelessness during their lifetime, with up to 35% of these individuals suffering from severe mental illness (p.393). While the review by Coldwell and Bender (2007), the sole review focusing on homeless individuals, holds many merits, the focus of the review is exclusively meta-analytic, stopping short of explaining how and why ACT programs support (or fail to support) housing tenure among homeless individuals. A review aiming to better understand how and in what way ACT programs support housing tenure among homeless individuals with mental illness holds potential promise to inform future program planning and implementation of ACT programs tailored to this notoriously difficult to treat subpopulation.

Taken collectively, then, the time is ripe for a review of ACT that (1) provides a better understanding of the ACT program components that promote positive housing tenure; (2) examines any unexplained variation across primary studies; and (3) fully integrates a broader range of studies—covering implementation as well as effectiveness—on ACT programs for homeless individuals. The following application of meta-modeling on ACT speaks directly to these aims.

CHAPTER 5. FRAMING THE META-MODELING SYNTHESIS

The following Chapters 5 through 8 illustrate a meta-modeling application on Assertive Community Treatment programs for homeless individuals with severe mental illness. The motivation for the application is two-fold in that it both (1) serves to illustrate a practical application of meta-modeling and at the same time (2) reveals insights about the program effectiveness of ACT programs. Structured according to the main procedural steps of meta-modeling, the chapters offer step-by-step procedural guidance on how to apply meta-modeling in the context of social programs.

In the present chapter, the first three steps of the meta-modeling approach, which serve to frame and develop the empirical grounding for the synthesis, are demonstrated. In line with traditional synthesis approaches, the primary purpose of these steps is to frame and ensure the systematic and transparent development of a comprehensive empirical grounding for the subsequent analyses. For each step, a set of methodological reflections is provided.

Step 1. Define the Scope (Research Question)

The first step of meta-modeling is to develop a research question. For the present purposes, the overarching research question guiding the review was two-fold:

- (1) *To what extent do Assertive Community Treatment (ACT) programs, as compared with standard care, increase housing retention among homeless individuals with severe mental illness?*
- (2) *What, if any, ACT components drive the increased housing tenure among homeless individuals with severe mental illness?*

The first question is a classic effectiveness question, focusing on the comparative advantage of ACT programs, as compared with standard care. The primary purpose of the question is to assess the effectiveness of ACT in terms of supporting homeless individuals with severe mental illness obtain and maintain housing. Answering this question is a necessary stepping-stone for establishing a causal connection between ACT programs and housing tenure.

The second question reaches beyond ‘whether ACT works’ in that it calls for information about how and in what way ACT programs support housing tenure among homeless individuals with severe mental illness. Answering this question results in a better understanding of how and in what way ACT programs work (or fail to work). The primary purpose is to better understand the programmatic components that are critical for effective ACT programs.

Following the PICO guidelines, the research question specifies the population (e.g., *homeless individuals with severe mental illness*), the intervention (e.g., *Assertive Community Treatment*), the comparison (e.g., *standard care*), and the desired outcome (e.g., housing tenure). The question is exploratory in the sense that it does not have a specific hypothesis about the relative performance of ACT in relation to comparable standard care programs, or about specific programmatic components driving the program performance.

Methodological reflections

Several observations on formulating questions for a meta-modeling synthesis are worth sharing. First and foremost, in the present context, the formulation of the research questions was informed by the scoping review of 14 published reviews of ACT programs (see Chapter 4 above). The scoping review not only served to identify a gap within the existing set of ACT reviews but also supported the specification of the PICO criteria and a better understanding of the volume and

major strands of research on ACT programs. As such, the scoping review provided a strong foundation on which to develop a more useful and focused synthesis.

Secondly, and even when grounded on a scoping review of existing reviews, the optimal and feasible level of specificity in the question formulation is difficult to determine. If framing the question, and in effect the synthesis, too broadly, the possibility of introducing program variations (and large amounts of heterogeneity) in the analyses is increased (see Chapter 9 for a more detailed discussion of the implications of heterogeneity). Conversely, framing the questions, and in effect the synthesis, too narrowly runs the risk of resulting in too few qualified studies to support a meta-modeling synthesis. There is no clear cut solution to this.

Third and finally, the two-fold formulation of the research question represents a useful structure for highlighting the fundamental difference between determining the net-effect of a program and explaining how the program generates same said net-effect. Whereas traditional meta-analyses, if implemented well, may serve to answer the first question (does ACT work?), and in effect establish a causal connection between ACT and housing tenure, for instance, these classic meta-analyses will more likely fall short of adequately answering the second question, that is, identifying the program components driving the observed effects (how does ACT work?). The second question simply demands a different type of information and in effect a different type of data and analytical approach. In the present context, and as is illustrated below, meta-modeling includes findings from implementation studies and rely on Qualitative Comparative Analysis to address the second research question (see Chapter 8). The position held by the author of the present dissertation is that each question is equally valid and mutually called for when establishing and explaining causal relationships between programs and outcomes.

Step 2. Search and Retrieve Relevant Studies

In logical extension of formulating the research questions, the second step of the meta-modeling approach is to systematically search and retrieve relevant studies.

For the present project, a structured multi-pronged sampling strategy was deployed to identify candidate studies. More specifically, primary studies were identified through an electronic literature search using the Web of Science (1970-2016), PsycINFO (1974-2016), PubMed (1950-2016), Sociological Abstracts (1952-2016), ProQuest Dissertation and Theses Global (2016), CINAHL (1982-2016), Social Services Abstracts (1995-2016), and the Cochrane Central Register of Controlled Trials databases (2016). Collectively, these databases provide sufficient coverage of the major publication outlets for studies in the social and health sciences, including psychology, public health, and social work. With the aim of identifying relevant studies, the search term “Assertive Community Treatment” in conjunction with “Homeless*” were used. Only manuscripts in English were included. No restrictions on publication date, study design, or study location were used.

In addition to the electronic search, the reference lists of 36 literature reviews (identified in the initial scoping review) were manually examined for relevant studies. Of these, 22 research reviews covered case management in general and 14 reviews were on ACT more specifically (see Table 4.1 above). The manual search of these reviews served double-duty by allowing for the reviewer to crosscheck references with the electronic search and by gauging whether a point of saturation had been reached, that is, whether the manual search reached a point where no new relevant studies were revealed. Figure 5.1 in Chapter 5 provides an overview of the search and retrieval process, including the number of references derived from the electronic and manual searches. As the figure shows, the electronic search identified 278 unique manuscripts. The

subsequent manual review of literature reviews resulted in 246 additional manuscripts. The total 524 unique manuscripts were organized in an excel spreadsheet, detailing for each manuscript the author name(s), manuscript title, type of manuscript (e.g., journal article, book chapter, dissertation), journal name, abstract (in full), and publication year. The list was formatted for consistency and double-checked for duplicate records.

Methodological reflections

Many important lessons learned emerged from the search and retrieval of relevant ACT studies, of which two are particularly worth noting. The use of the label ‘Assertive Community Treatment’ as a search term is deceptively simple and not without its limitations. This is in part because the ACT label may be used for studies that upon further inspection prove not to be ACT programs (false positives) or may not be used by programs that are in fact ACT programs (false negatives). The inclusion of false positives, programs labeled as ACT without being ACT programs, is a lesser concern in that it can be remedied by careful assessment of the program description in the manuscript. However, the potential issue of false negatives, ACT programs not labeled as such, is more difficult to remedy. In the present context, the manual reference search in existing reviews, many of which reached beyond ACT and covered intensive case management more broadly, served as a way to identify potential false negatives and remedy this concern.

In a similar vein, another limitation of the search and retrieval of ACT studies pertains to the deliberate exclusion of gray literature (i.e., unpublished studies). Because the primary aim of the present project is methodological—to apply and examine a new approach for mixed methods synthesis—the decision not to include gray literature is considered acceptable. That being said,

the exclusion of gray literature introduces publication bias to the meta-analytic findings of the synthesis. To remedy this concern, the presence and potential influence of publication bias is examined empirically, as part of the effectiveness meta-summary (see Chapter 7).

Step 3. Conduct a Relevance Appraisal of the Studies

The third step of the meta-modeling approach is to conduct a relevance appraisal of the manuscripts identified in the electronic and manual searches. In the context of the present review, the relevance appraisal was structured according to two stages. In the first stage, each abstract was carefully read to identify manuscripts that were clearly out of scope (e.g., a method note on measuring risk factors for hospitalization), unpublished reports (e.g., government reports or grant proposals), or irrelevant publication types (e.g., commentaries or opinion pieces). A total of 102 manuscripts were excluded on the basis of this initial abstract screening.

In the second stage of the relevance appraisal, the remaining 422 manuscripts were subsequently read in full and considered for inclusion in either the effectiveness meta-summary or the implementation meta-summary.

For the effectiveness meta-summary, the inclusion criteria were that studies should (1) be based on experimental or quasi-experimental designs (i.e., containing a control/comparison group), (2) compare Assertive Community Treatment to standard case management care (inpatient/outpatient), (3) provide findings specifically to homeless individuals with severe mental illness, (4) and include housing tenure (however defined) as an outcome measure. A total of 20 manuscripts met these criteria and were included for further analysis as part of the effectiveness meta-summary.

For the implementation meta-summary, the inclusion criteria were that studies should (1) pertain to Assertive Community Treatment program and (2) focus on fidelity of implementation (e.g., studies on critical ingredients, fidelity measures, or implementation studies). No criteria related to study design, subpopulation, or outcomes were considered (for reasons discussed below). A total of 26 manuscripts were included for further analysis as part of the implementation meta-summary.

The remaining 376 manuscripts, fulfilling neither the inclusion criteria for the effectiveness nor the implementation meta-summaries, were excluded from further consideration and analysis. These were studies that while interesting in their own right served no purpose in the specific context of meta-modeling ACT programs on housing tenure for homeless individuals with severe mental illness. Figure 5.1 provides a flowchart of the manuscript search, retrieval, and relevance appraisal.

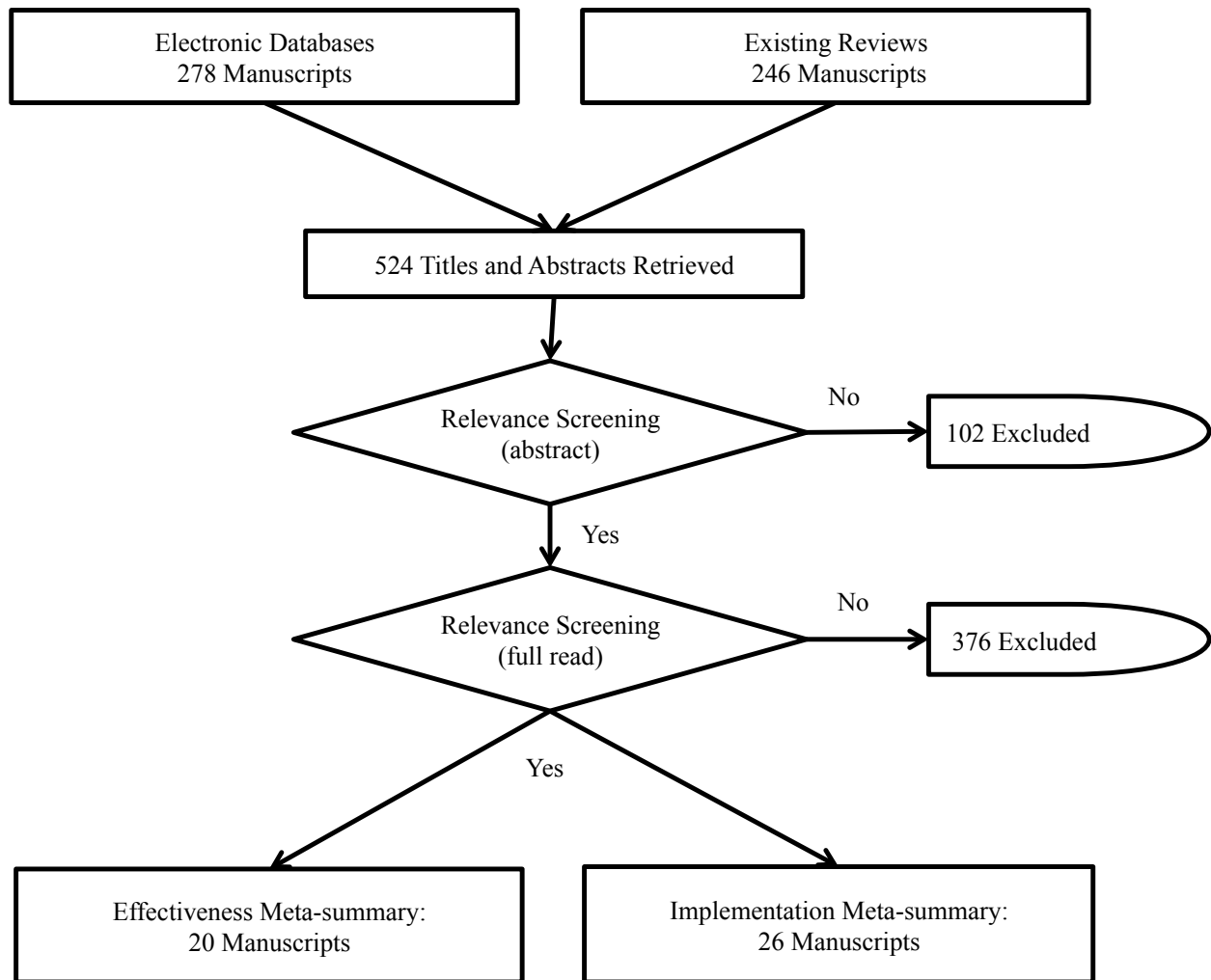


Figure 5.1 Flow chart of manuscript search and relevance appraisal

Methodological reflections

A few methodological observations are worth highlighting in regards to the relevance appraisal. First and foremost, the initial abstract-based relevance appraisal should be carried out with great care and with a lean towards erring on the side of including any manuscripts that appear even remotely relevant for further analysis. Aiming for a comprehensive review, the potential loss caused by excluding a relevant manuscript supersedes that of including an irrelevant manuscript for a subsequent full-read appraisal. The latter only comes at the cost of reviewer time and

resources spend on reading the manuscript in full. Conversely, discarding a relevant manuscript on the basis of an abstract-review could potentially influence the results and conclusions of the synthesis.

Another important observation pertains to the identification of Assertive Community Treatment programs in the included manuscripts on the basis of manuscript authors' labeling of the program under study as Assertive Community Treatment, ACT, or PACT. This identification sometimes appears frail, especially since the descriptions of these programs often vary markedly across the manuscripts and at times, to make matters more complex, even suggest noticeable variations across the programs presented. Accordingly, the confidence with which programs were identified as ACT varied somewhat across the included studies.

To make matters more complex, and as noted by Monroe-DeVita et al. (2011), among others, some of the defining features of ACT, such as in-vivo service delivery, have increasingly been assimilated by other care management programs, why the distinction between ACT and the broader array of intensive case management programs, over time, have become increasingly blurred. In a related vein, the descriptions of the standard treatment programs were often awarded even less attention, resulting in uncertainty about the nature of these standard care program and in effect the comparative advantage of ACT (comparative according to what?). The issue of blanket labeling and minimal descriptions of standard treatment programs is admittedly nothing new in literature on research synthesis. However, nowhere is this more problematic than in the context of specifying critical ingredients on the basis of synthesizing findings from existing studies. Accordingly, I will consider these related issues, as well as their specific implications in relation to ACT and meta-modeling, as part of the discussion in Chapter 9.

CHAPTER 6. THE IMPLEMENTATION META-SUMMARY

Step 4: Summarize the Critical Components in the Implementation Studies

The overarching purpose of the implementation meta-summary is to generate a set of hypotheses about the specific programmatic components, or configurations thereof, which bring about positive outcomes. Towards this end, the development of the implementation meta-summary is structured around two coding cycles: (1) one cycle of concept coding (within studies) and (2) one cycle of meta-coding (across studies). On the basis of these sequential coding cycles, the resultant meta-summary specifies a set of ‘critical’ program components and a narrative description of how these are hypothesized to promote positive outcomes.

The first coding cycle: Conceptual coding

Informed by Miles, Huberman, and Saldaña’s (2014) “conceptual coding”, and as described in Chapter 3 above (see section on the implementation meta-summary), the first cycle of coding aims to identify and extract the core program components of ACT. The core idea of this coding is that these conceptual models reflect an underlying theory about the most salient ACT program components driving the desired outcomes.

In the present review, a total of 26 manuscripts pertained to conceptual and empirical examinations of the core components of ACT. Based on a careful reading of these, five manuscripts that developed and empirically examined specific conceptualizations of the core components comprising ACT were identified (see Table 6.1). These five models, which were commonly applied and cited in the broader pool of implementation studies, were developed sequentially, each successive model building on the previous. The presentation of the models

was in the form of manuscripts detailing the conceptual development and empirical testing of corresponding fidelity measures, resulting in both specific and operationalized definitions of core ACT components. As such, they were particularly salient for the purposes of identifying potential critical ingredients of ACT programs.

The five manuscripts revealed several approaches for developing ACT fidelity measures, including existing literature (Teague et al., 1995), existing literature and expert surveys (McGrew & Bond, 1995; McGrew et al., 1994), or some combination thereof (Monroe-DeVita et al., 2011; Teague et al., 1998). The resultant fidelity measures were often pilot-tested on a sample of ACT programs, supporting an empirical examination of their measurement properties (e.g., factor structure, reliability) and practical applicability and utility (e.g., respondent burden, ability to distinguish between ACT program variations). As the table indicates, the measures varied in both their main dimensions and specific items; although, further analyses revealed noticeable overlap in regards to the latter. Finally, the measures typically relied on a five- or seven-point numeric scale, with different variants of text anchors.

Table 6.1 Published fidelity measures of Assertive Community Treatment

	McGrew et al. (1994)	McGrew & Bond (1995)	Teague et al. (1995)	Teague et al. (1998)	Monroe- DeVita et al. (2011)
Name	The Index of Fidelity to ACT (IF- ACT)	The Critical Components of ACT Interview (CCTI)	No name	The Dartmouth ACT Scale (DACTS)	The Tool for Measurement of Assertive Community Treatment (TMACT)
Dimensions	Three dimensions	Nine dimensions	Two dimensions	Three dimensions	Six dimensions
Items	17 items	54 items	13 items	26 items	47 items
Scale	Seven-point scale with anchors for each point.	Seven-point scale with anchors for each point.	Five-point scale with anchors for each point	Five-point scale with anchors for each point.	Five-point scale with anchors for each point.
Source	Existing literature and expert opinion	Existing literature and expert opinion	Existing literature	Existing literature and expert opinion	Existing literature (especially DACTS measure) and expert opinion

To advance the first coding cycle, the items comprising the five models were listed and organized in a content summary table (adapted from Miles et al., 2014). For illustrative purposes, an abridged version of the content summary table is provided in Table 6.2. As the table indicates, while there is a great deal of overlap across the five measures, significant variation emerges in regard to the specific definition of individual components. Suitably, the content summary table facilitated a more focused examination of these variations across the five fidelity measures. By having the specific components grouped according to larger, more comprehensive categories, the variability in how different fidelity measures defined specific components was easier to assess.

The distinctions and similarities were easier to identify. To illustrate, while McGrew and Bond (1995) emphasize the presence of a psychiatrist on the ACT team, Teague et al. (1998) specify at least one full-time psychiatrist per 100 clients, and Monroe-DeVita et al. (2011) consider both the presence and the role of the psychiatrist in treatment and within the team.

In addition to facilitating a better understanding of the program component variations, the total list of components in the content summary table can be viewed as a comprehensive list of candidate program components for further analysis. In the present analysis, this list amounts to 61 candidate program components, an unfeasible number of components for further empirical consideration. Towards identifying the most salient critical components, the second level of meta-coding is called for.

Table 6.2 Content summary table for ACT components (abridged)

Core Component	McGrew et al. (1994)	McGrew & Bond (1995)	Teague et al. (1995)	Teague et al. (1998)	Monroe-DeVita et al. (2011)
Psychiatrist on team	Psychiatrist on team	Psychiatrist on team	No mention	Psychiatrist on staff (at least one full-time psychiatrist per 100 clients)	-Psychiatric care provider on team -Role of psychiatric care provider (in treatment) -Role of psychiatric care provider (within team)
Caseload	Client/staff ratio	Small client/staff ratio	Small caseload	Small caseload (client/provider ratio of 10:1)	Low ratio of consumers to staff
Team approach	Shared caseloads	Team approach -Shared caseloads for treatment planning -Shared caseloads for service provision	Team approach	Team approach (provider group functions as a team rather than individual practitioners; clinicians know and work with all clients)	Team approach

The second coding cycle: Meta coding.

The purpose of the second coding cycle is to develop a refined list of the most salient critical components. This requires the analytical level to shift from coding within studies to coding and

analysis across studies. In its practical application, the cross-study coding involves two analytical steps: (1) identifying the most salient components across the subset of fidelity studies (cross-study analysis); and (2) developing a description of the underlying reasoning for each of the selected components (component specification).

Informed and structured by the initial content summary table, the cross-study analysis led to the identification of five core ACT components: (1) low client/staff ratio, (2) 24-hours crisis services, (3) team approach, (4) in-vivo service delivery, and (5) time-unlimited services. These are summarized in Table 6.3 below. The five components were consistent across all five ACT fidelity measures and serve to distinguish ACT from the more broader category of intensive case management. As such, they can reasonably be viewed as core components of ACT programs.

Towards developing an operational description of each of the five components, to be used in subsequent integrated synthesis, their underlying mechanisms and full implementation were formulated (see Table 6.3 below). These component specifications were derived from the DACTS and TMACT protocols, the two most widely applied ACT fidelity measures (Monroe-DeVita et al., 2011; Teague et al., 1998).

Table 6.3 Five core components of Assertive Community Treatment

Core Component	Mechanism	Full Implementation
Low Client/Staff Ratio (CSR)	Low client to staff ratio supports increased intensity and individualization of services.	Client to staff ratio is equal or lower than 10:1.
24-hour Services	Provision of crisis services 24 hours a day, seven days of week, supports immediate and continuous care.	Crisis services are provided 24 hours per day, seven days a week.
Team Approach	The team approach ensures immediate and direct services.	Caseload is shared by entire ACT team.
In-vivo Delivery	Providing services where clients live, work, and interact with other people focuses the care on the client's immediate needs and personal goals, enhance client engagement, and allows for better transfer of skills.	Consumer contact primarily in community.
Time-Unlimited Treatment	Time-unlimited services encourage the development of stable, ongoing therapeutic relationships.	Program offers time-unlimited services for all clients.

Table note: Mechanism descriptions adopted from (Monroe-DeVita et al., 2011; Teague et al., 1998).

Methodological reflections

A central—yet all too often neglected—question in mixed methods synthesis literature revolves around what constitutes data when coding qualitative primary studies. Is the coding directed at author findings, raw quotes provided by the author in support of these, or a combination of both. The meta-modeling approach, by focusing the analytical lens on the empirically validated fidelity measures, lessens the severity of this challenge. Moreover, the meta-modeling focus on these fidelity measures also capitalizes on the operational and empirically supported nature of these measures, which in turn support more specific and empirically-driven hypotheses about the critical program components.

One aspect of the fidelity measures that holds less application in the context of meta-level synthesis is that of rating scales, that is, the operational definition of how to rate the degree of adherence to each component. This is because the developed fidelity measures are based on applications where information from clinical records, administrative staff logs, and even individual interviews and ethnographic observations may support the fidelity ratings. Accordingly, the scales and scoring schemes are reliant on these types of information sources, which for all practical purposes are unrealistic in the context of a systematic review. The sole data source for the latter is the program description provided in the published manuscript. If individual studies were to provide fidelity scores (from comparable fidelity measures) these could be included and examined in the meta-analysis as a moderator of the combined effect size.

Another observation worth sharing is that developing an implementation meta-summary is not just a preparatory step of data extraction and organization for the purposes of further analyses. This is because each coding cycle represents an interpretative act. As just one example, and while deceptively simple in its operation, conceptual coding is highly interpretive. This is in part because the critical components may be labeled and described in different ways across the included studies. As such, careful consideration should be awarded the definition and description of each component identified, ensuring that the meaning of the critical component does not shift substantively across the measures. In a similar vein, the simple act of ordering the initial components according to larger categories involves decisions and interpretations on the meaning of the components mentioned in each manuscripts, many of which are open to interpretation (simply consider the many potential interpretations of “assertive engagement”).

Finally, and in a related vein, each step of the coding reveals numerous insights about the program under review. In the present context, reading and coding the implementation

manuscripts revealed numerous generations of program variations, shifts in core components over time, and changing definitions and meanings of these. These insights provide a foundation for the later interpretation of causal recipes (see Chapter 8), may serve to inform recommendations for further research, and even hold value for future program planning. In this way, the findings of the implementation meta-summary hold value independently of subsequent analyses.

CHAPTER 7. THE EFFECTIVENESS META-SUMMARY

Step 5: Summarize the Effect Sizes in the Effectiveness Studies

The aim of the effectiveness meta-summary is to examine the effect of the intervention through a meta-analysis of the experimental and quasi-experimental studies identified. This serves both to gauge the overall effect of the program under review as well as to calculate individual effect sizes to be included in the final integrated meta-model synthesis. In line with the development of the implementation meta-summary, there are two successive steps in developing the effectiveness meta-summary: Coding and estimating effect sizes (within each study) and summarizing these in the form of a meta-analysis (across studies).

Coding and estimating effect sizes (within study)

The first step in the coding of the manuscripts revolved around coding the basic characteristics of the studies described in the 20 manuscripts. These are provided in Table 7.1 below. As the table shows, study variations included population characteristics (e.g., homeless individuals versus individuals at risk of homelessness), treatment programs (e.g., ACT versus ACT plus housing support), and control programs (e.g., standard treatment versus standard treatment plus housing support). The variations in treatment programs, that is, ACT versus ACT plus housing provision, deserve special attention. As the table reveals, the most recent studies of ACT programs involve ACT programs with integrated housing provision. These hybrid programs emerge from the general push towards supported housing in general and Housing First more specifically. In the present context, especially given the focus on housing tenure as primary outcome, this programmatic variation is important because the direct comparison (or integration) of effect sizes

from ACT programs with and without provision of housing is inappropriate; the programs are fundamentally different programs.

Table 7.1 Study Characteristics (by study design)

Study	Location	Population	Treatment	Control
Experimental:				
Morse et al. (1992)	St. Louis	Homeless mentally ill people	ACT	Standard outpatient treatment/drop-in center
Korr & Joseph (1995)	Chicago	Homeless (or at risk of homelessness)	ACT (plus housing)	Standard outpatient treatment/drop-in center
Lehman et al. (1995)	Baltimore	Homeless persons with severe and persistent mental illness	ACT (plus consumer advocates & family outreach)	Standard treatment
Lehman (1997)	Baltimore	Homeless persons with severe and persistent mental illness	ACT (plus consumer advocates & family outreach)	Standard treatment
Morse et al. (1997)	St. Louis	Homeless (or at risk for homelessness) with severe mental illness.	ACT/ACT with community workers	Brokered case management
Wolff et al. (1997)	St. Louis	Homeless (or at risk for homelessness) with severe mental illness.	ACT/ACT with community workers	Brokered case management
Holter (1998)	New York	Homeless mentally ill men	ACT (Critical Time Intervention)	Standard treatment
Lehman et al. (1999)	Baltimore	Homeless persons with severe and persistent mental illness	ACT (plus consumer advocates & family outreach)	Standard treatment
Clarke et al. (2000)	Portland	Chronically mentally ill (homeless in the	ACT (+/- community workers)	Standard treatment

Rosenheck et al. (2003)	San Francisco	past six months or at risk for homelessness) Homeless veterans with psychiatric and/or substance abuse disorders	ACT	Standard treatment (VA services)
Kenny et al. (2004)	St. Louis	Homeless (or at risk for homelessness) with severe mental illness.	ACT/ACT with community workers	Brokered case management
Morse et al. (2006)	St. Louis	Homeless with co-occurring severe mental illness and substance use disorder (dual diagnosis)	ACT/ Integrated ACT	Standard treatment
Fletcher et al. (2008)	St. Louis	Homeless with co-occurring severe mental illness and substance use disorder (dual diagnosis)	ACT/ Integrated ACT	Standard treatment
Quasi- experimental:				
Drake et al. (1997)	Washington, DC.	Homeless with co-occurring severe mental illness and substance use disorder (dual diagnosis)	ACT	Standard treatment
Morris & Warnock (2001)	Tulsa	Homeless with severe mental illness	ACT Mobile Outreach	Standard Mobile Outreach
Morse et al. (2008)	St. Louis	Homeless with co-occurring severe mental illness and substance use disorder (dual diagnosis)	Integrated ACT	Standard Treatment

Young et al. (2009)	Hillsborough	Homeless individuals with complex behavioral health needs	ACT (plus housing)	Standard treatment
Mares et al. (2011)	Chattanooga	Homeless individuals with mental health, substance abuse or medical needs.	ACT (plus housing)	Comprehensive, Continuous, Integrated System of Care
Somers et al. (2015)	Vancouver	Homeless and mentally ill	ACT (plus Housing First)	Standard treatment plus congregate housing
Clark et al. (2016)	Tampa	Homeless with long histories of hospitalization for serious mental illness	ACT (plus supported housing)	Critical Time Intervention (plus supported housing)

The second step of the synthesis was to code and extract data from each study to support the calculation of standardized effect sizes. This abstraction included data on sample sizes for treatment and control groups (after attrition), mean statistics and corresponding standard errors on housing tenure outcomes, and/or proportions or percentages for housing tenure outcomes, or test statistics from reported significance tests (e.g., chi-square value, t-test value, and/or corresponding *p*-values). For the purpose of creating linearly equivalent outcome measures, basic data transformations were made for some of the studies. For instance, days homeless during a specified time period were transformed into days stably housed (by subtracting the former from the total days of the time period).

On the basis of the retrieved information, effect sizes were calculated for each study using the Practical Meta-Analysis Effect Size Calculator (Lipsey and Wilson, 2001). The effect size measure calculated in the present review was Cohen's *d*, the standardized mean difference

statistic (though, any other standardized effect size measure would be acceptable). Because of the relative small samples in some of the studies, all the individual effect sizes were adjusted for small sample bias, using the Hedges g correction (Hedges and Olkin, 1985).

Finally, those studies that were reported on in multiple manuscripts were combined and treated as one study for the purposes of the meta-summary. This resulted in 15 unique ACT studies. The resultant effect size estimates for each of the 15 unique studies are provided in Table 7.2 below.

A couple of observations are noteworthy. First and foremost, the included studies revealed noticeable variations in outcome measures. Whereas some studies defined housing tenure as mean number of days housed/homeless (within a specified time period), other studies defined it as proportion of participants in housing or experiencing homelessness. Secondly, the sample sizes varied markedly across studies, ranging from a sample size of 39 to a sample size of 278 individuals. Finally, the timeline for outcome measurement ranged from 6 to 36 months across the studies.

Parallel to these study variations, and possibly reflecting these, the estimated effect sizes also varied noticeably across the studies. Whereas most studies indicated small to moderate positive effect sizes (0.12 to .50), seemingly favoring ACT over standard treatment, several of these studies displayed confidence intervals with values below zero, suggesting null effect or even a negative effect size range. These patterns can also be examined in the corresponding forest plot (provided in Figure 7.1 below). These program variations, and the resultant ambiguous effect size patterns, are not uncommon. After all, the very motivation for research syntheses emerges from the variation encountered across studies. Accordingly, the nature and implications of these variations are considered as part of the meta-analysis that follows.

Table 7.2 Individual effect sizes (Hedges g) for effectiveness studies (by program type)

Study design	N ^a	Outcome Measure	Timeline	Effect Size (g)	95% CI
ACT:					
Morse-St Louis (1992)	72	Number of days homeless (one month prior)	12 months	0.29	-0.18 to 0.76
Lehman-Baltimore	125	Number of days homeless (one year)	12 months	0.50	0.14 to 0.86
Drake-Washington	187	Number of days housed (two months prior)	18 months	0.46	0.12 to 0.79
Morse-St. Louis (1997)	135	Number of days housed (one month prior)	18 months	0.39	0.05 to 0.74
Holter-New York	89	Number of days homeless	18 months	0.29	-0.13 to 0.72
Clarke-Portland	163	Proportion experiencing at least one homeless episode	24 months	0.12	-0.22 to 0.46
Morris	50	Number of days homeless (6 months prior)	6 months	1.13	0.52 to 1.74
Rosenheck-San Francisco	278	Number of days in stable housing	36 months	0.14	-0.12 to 0.39
Morse-St. Louis (2006)	148	Number of days housed	24 months	0.22	-0.13 to 0.57
Morse-St. Louis (2008)	128	Number of days housed	18 months	0.34	-0.03 to 0.70
Young-Hillsborough	81	Proportion housed	6 months	-0.37	-0.83 to 0.08
ACT (with housing):					
Korr-Chicago	95	Proportion housed	6 months	0.88	0.45 to 1.31
Mares-Chattanooga	39	Proportion housed	24 months	1.92	1.01 to 2.84
Somers-Vancouver	197	Percentage of time in stable housing	12 months	0.01	-0.27 to 0.29
Clark-Tampa	185	Proportion housed	6 months	0.34	0.04 to 0.64

Table note: ^a total sample size before attrition.

Meta-analysis: Summarizing effect sizes (across studies)

On the basis of the individual effect sizes, a meta-analysis was subsequently conducted using Meta-Essentials for Excel (Van Rhee, Suurmond, & Hak, 2015). Three steps comprised the meta-analysis: (1) Calculation and interpretation of combined effect sizes and heterogeneity statistics, (2) subgroup analysis (by research design and program type), and (3) publication bias analysis.

Following Lipsey and Wilson (2001), the combined effect size across all 15 ACT studies was estimated by fitting a random-effects model and using inverse variance weighting, whereby each study is weighted by the precision of their respective effect size estimate. The individual and combined effect sizes are provided in the forest plot presented as Figure 7.1 and Table 7.3.

In the forest plot, the effect sizes (and corresponding 95% confidence intervals) for the individual studies are presented in rows 1-15, with positive effect differences favoring ACT. A study-level difference of zero indicates no difference in housing tenure between ACT and standard treatment. The size of the marker for each individual effect size reflects the relative contribution (i.e., weight) of the effect size to the combined effect size for all the studies. The combined effect size, presented in row 16, is presented with a corresponding 95% confidence interval and a predication interval. The latter expresses the range within which future effect size estimates for ACT will fall.

As shown in Table 7.3, the combined effect size (across all 15 effectiveness studies) is 0.36 (95% CI: 0.13 to 0.60), suggesting a small to medium effect of ACT, as compared with standard treatment. Because the lower boundary of the 95% confidence interval is above zero, we may tentatively interpret the effect size as statistically significant ($p < 0.05$). Before drawing any firm conclusions, however, and especially given the program design variations identified

above, the amount of effect size variation demands our attention. As the forest plot indicates, the bandwidth of the prediction line, indicating the predicted bandwidth of effect sizes of future ACT programs, ranges from -0.29 (a small negative effect) to 1.01 (a large positive effect). From this perspective, the combined effect size estimate becomes less, if at all, meaningful because it hides more than it reveals. As a result, all we can conclude on the basis of this overarching effect size estimate is that the extent to which ACT works is likely more complex than initially anticipated.

As shown in Table 7.3, the unexplained variation among the ACT effect sizes was examined using I^2 statistic, which expresses the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies (Higgins et al., 2002). The estimated I^2 statistic of 0.70 indicates that 70% of the variation across the estimated effect sizes is due to study heterogeneity. This high level of heterogeneity implies that the studies on which the combined effect size is based cannot reasonably be viewed as a sample of studies emerging from the same, homogenous population of studies. There are genuine differences underlying the results of the studies, why estimating a combined effect size across all 15 studies is not meaningful. Advancing the analysis further, then, effect sizes for specified subgroups should be considered.

Table 7.3 Random-effects model (all 15 ACT studies)

	Hedges g	z-value	p-value	95% CI	Prediction interval	I^2
Combined effect size	0.36	3.29	0.001	0.13-0.60	-0.29-1.01	0.70

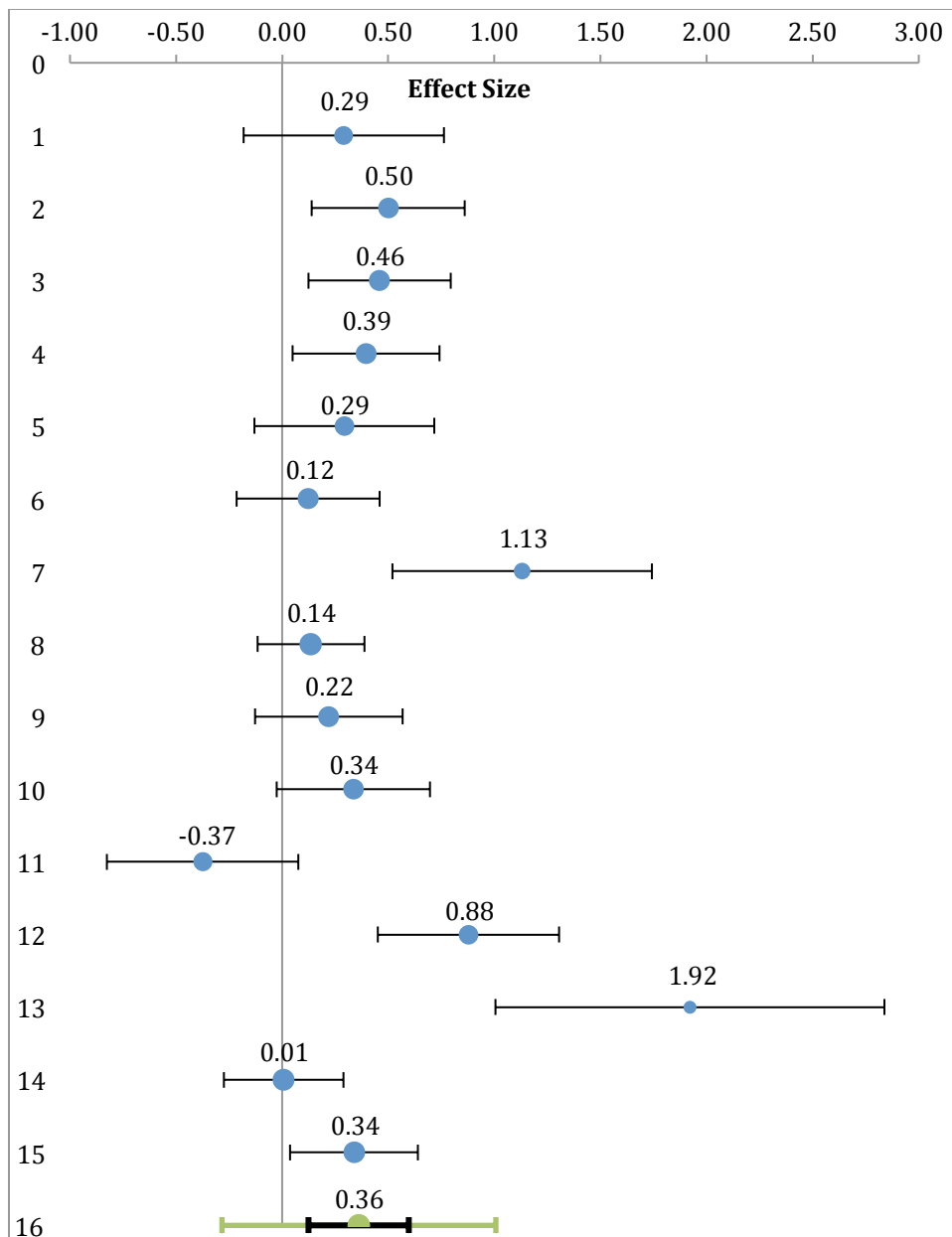


Figure 7.1. Forest plot of individual plus combined effect size (Hedges *g*)

Informed by the high amount of unexplained variation, and with the aim of establishing more homogenous and comparable subgroups, the 15 studies were grouped and analysed by research design (experimental and quasi-experimental) and program type (ACT and ACT plus housing provision) were conducted. Variations due to study designs are well documented in the synthesis

literature (Shadish & Ragsdale, 1996), with experimental designs often producing smaller and more accurate effect sizes. In the present context, the distinction between regular ACT programs and ACT programs awarding direct access to housing also seemed salient, especially given the focus on housing tenure as primary outcome. For these programs, by providing access to housing, the program provides the outcome directly, why comparisons with regular ACT programs seem unequivalent.

The categorization by research design and program type resulted in four subgroups: Experimental studies of ACT programs; quasi-experimental studies of ACT programs; Experimental studies of ACT programs with direct housing provision; and quasi-experimental studies of ACT programs with direct housing provision. The subgroup analyses revealed several noticeable patterns. As Table 7.4 shows, the combined effect size for experimental studies of ACT programs is markedly smaller ($g=0.26$), as compared with quasi-experimental studies of ACT programs ($g=0.36$), as well as experimental and quasi-experimental studies of ACT programs with housing support ($g=0.43$ and $g=1.04$, respectively).¹ Falling in line with the rationale for selecting study design and program type for subgroup analyses, these observed effect size variations are easily explainable, even to be expected.

Also noticeable is the variation around the effect sizes for the subgroups, which displays a distinct pattern, with noticeably larger confidence intervals for quasi-experiments and for programs combining ACT with housing provision, as compared with experimental studies of

¹ The seven experimental studies of ACT indicate decreasing effect sizes over time. This noticeable pattern could, at least in part, reflect the spillover and absorption of ACT principles in standard treatment programs. For instance, the use of assertive outreach was as part of the broader dissemination of ACT programs in the 1990s increasingly adopted by other care management approaches. As such, the spillover of ACT principles into standard treatment programs may have resulted in reduced effect sizes over time. This pattern and plausible explanation is of course not based on any formal analysis.

ACT. This pattern is clearly visible in Figure 7.2, displaying a forest plot the effect sizes estimates (and corresponding 95% confidence intervals) for each subgroup of studies.

Of equal importance, and of special interest in the present context, the I^2 statistic for the seven experimental studies of regular ACT programs of 0.00% provides strong evidence that these studies can be viewed as a sample of a larger homogenous population of ACT studies. In the words of Hak et al. (2016), “studies in this subgroup have produced an estimate of the same “true” effect size in a homogeneous population”. In contradistinction, the variation between studies is much higher for quasi-experimental studies of ACT programs ($I^2=82.47\%$), experimental studies of ACT plus housing programs ($I^2=91.21\%$), and quasi-experimental studies of ACT plus housing programs ($I^2=90.90\%$). As such, these subgroups cannot be viewed as a sample of a homogenous set of studies, and in effect do not support any conclusions about the effectiveness of ACT. For the purpose of determining the overall effectiveness of the ACT programs, the estimate of the seven experimental studies provides the most reliable and accurate combined effect size estimate.

Table 7.4 Summary effect and heterogeneity for ACT and ACT plus housing (by study design)

	Study	Effect size	CI Low	CI High	I ²
ACT (Experiment)	Morse-St. Louis (1992)	0.29	-0.18	0.76	
	Lehman-Baltimore	0.50	0.14	0.86	
	Morse-St. Louis (1997)	0.39	0.05	0.74	
	Holter-New York	0.29	-0.13	0.72	
	Clarke-Portland	0.12	-0.22	0.46	
	Rosenheck-San Francisco	0.14	-0.12	0.39	
	Morse-St. Louis (2006)	0.22	-0.13	0.57	
Combined ACT (experiment):		0.26	0.15	0.36	0.00%
ACT (Quasi-experiment)	Drake-Washington, DC	0.46	0.12	0.79	
	Morris-unknown	1.13	0.52	1.74	
	Morse-St. Louis (2008)	0.34	-0.03	0.70	
	Young-Hillsborough	-0.37	-0.83	0.08	
Combined ACT (quasi-experiment):		0.36	-0.22	0.95	82.47%
ACT-housing (Experiment)	Korr-Chicago	0.88	0.45	1.31	
	Somers-Vancouver	0.01	-0.27	0.29	
Combined ACT-housing (experiment):		0.43	-0.43	1.28	91.21%
ACT-Housing (Quasi-experiment)	Mares-Chattanooga	1.92	1.01	2.84	
	Clark-Tampa	0.34	0.04	0.64	
Combined ACT-housing (quasi-experiment):		1.04	-0.48	2.63	90.90%

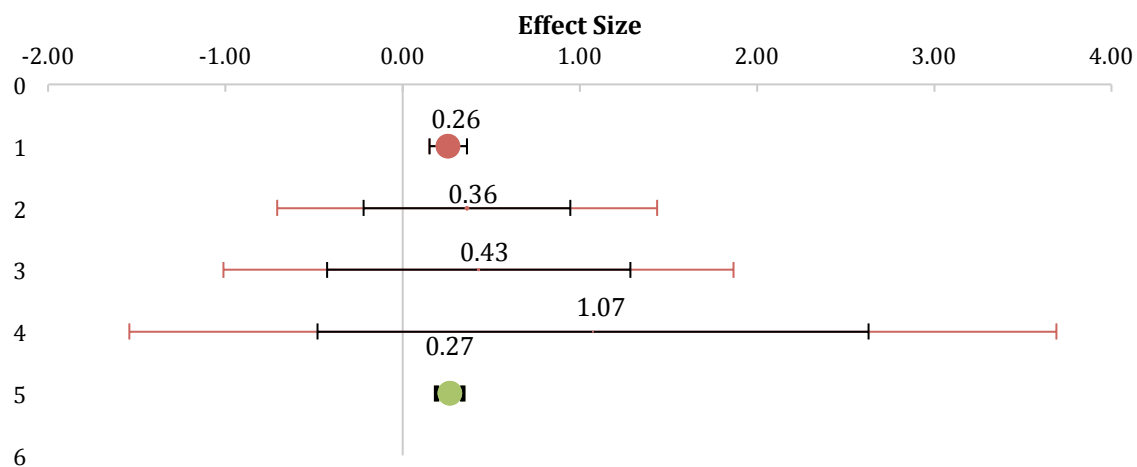


Figure 7.2 Forest plot of effect size estimates (by subgroup)

Informed by the subgroup analyses, a random-effects model was fitted on the seven experimental studies (comprising a homogenous subset of studies). The results are provided in Table 7.5. On the basis of these seven studies, the combined effect size is 0.26 (95% CI: 0.12 to 0.39), suggesting a small statistically significant effect of ACT on housing tenure, as compared with standard treatment ($p < .000$). Also worth noting, the prediction interval indicates that effect sizes of future ACT studies will fall within 0.12-0.39 (which is in the small to moderate effect size range). Finally, the I^2 statistic of 0.00 indicates a homogenous set of studies.

Table 7.5 Random-effects model (seven experimental studies on ACT)

	Hedges g	z-value	p -value	95% CI	Prediction interval	I^2
Combined effect size	0.26	4.76	0.000	0.12-0.39	0.12-0.39	0.00

Third and finally, the presence and influence of publication bias was examined. Because analyses of publication bias assume that the domain is homogeneous, the following analyses are based on the seven experimental studies of regular ACT programs. The funnel plot (shown in Figure 7.3) provides a visual summary of the effect sizes, plotting the effect size for each study against the corresponding standard error (basically study size). If the bias is limited, the distribution of effect sizes should appear symmetric and funnel shaped (as indicated by the red lines). As neatly articulated by Sutton (2009),

The premise for this appearance is that effect sizes should be evenly distributed (that is, symmetric) around the underlying true effect size, with more variability in the smaller studies than the larger ones because of the greater influence of sampling error (that is, producing a narrowing funnel shape as study precision increases) (p.437).

In the present analysis, the funnel plot included a trim-and-fill procedure, whereby smaller

studies causing asymmetry are removed ('trimmed'), a new center-line for the funnel plot is estimated, the removed studies and missing counterparts are replaced ('filled'), and an adjusted combined effect size is estimated.

As shown in Figure 7.3, and despite the relatively small number of studies, the individual effect sizes with similar standard errors do appear more or less symmetrically distributed around the combined effect size. Reinforcing this conclusion, the trim and fill procedure did not add any data points to achieve symmetry and the unadjusted and adjusted combined effect sizes are identical ($g=0.26$). In extension of the findings from the funnel plot, and because the subset of studies is homogenous, Rosenthal's fail-safe N was calculated, expressing the number of additional studies (in which the effect size is zero) that would be needed to result in the combined effect size to be statistically insignificant (Rosenthal, 1979). The estimated fail-safe N was 35 studies, suggesting fairly robust results.

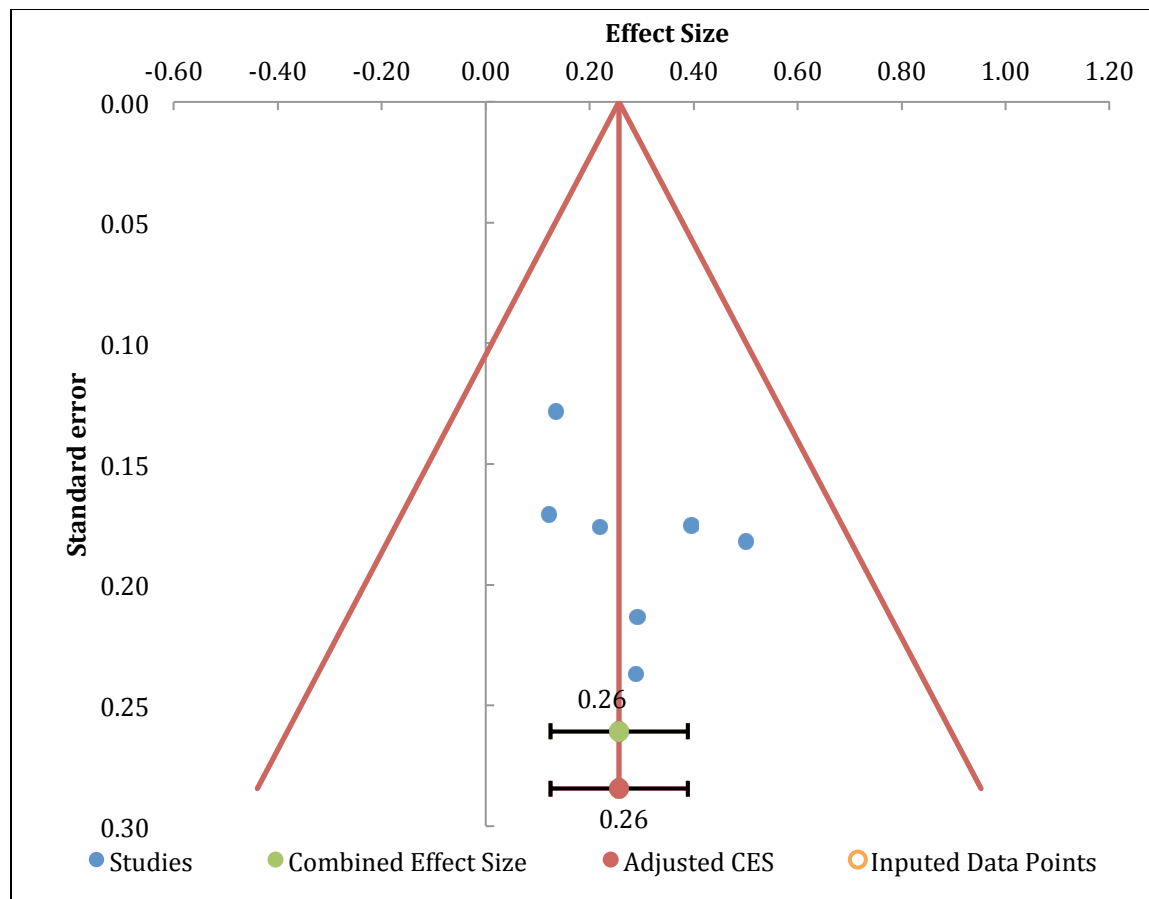


Figure 7.3 Funnel plot (with trim-and-fill)

To intermittently conclude, homeless individuals with mental illness demonstrate greater improvement in housing tenure when participating in ACT programs, as compared with similar individuals in standard treatment programs. In experimental studies, the combined effect size of ACT, as compared with treatment as usual, is statistically significant but small ($g=0.26$, 95% CI: 0.15 to 0.36). The influence of publication bias is minimal. Advancing the analysis further, the remaining variation in effect sizes across the seven experimental studies should be examined.

Methodological Reflections

The development of the effectiveness meta-summary yielded several noteworthy lessons learned about analyzing and synthesizing effectiveness studies. First and foremost, the amount of variability across the included studies posed a significant challenge for the purposes of estimating an accurate and reliable combined effect size. The marked variability in program variations, study designs, population characteristics, and operationalizations of the housing tenure outcome, is likely reflective of the broad dissemination and diverse application of the ACT model. As such, it reflects the many different purposes and applications of a model that while well-defined still encourages local adaptation.

However, in the context of a research synthesis, and especially within a meta-analytic framework, the variability made it difficult, almost impossible, to determine a single combined effect size estimate for ACT programs generally speaking. Instead, and this is not raised as a critique, the analytical emphasis of the meta-summary shifted towards examining dispersion of effect sizes for subgroups of studies. As a result, however, the number of studies contributing to the effect size estimate was dramatically reduced to seven experimental studies. The broader implications of large amounts of unexplained variations for the role and purpose of meta-analytic summaries will be considered as part of the discussion in Chapter 9.

Another lesson learned, and one that emerges from the variability discussed above, relates to the importance of maintaining a distinction between confidence and prediction intervals in meta-analyses. Whereas the former is quite commonly calculated and presented as part of meta-analytic summaries, the latter is rarely discussed and when considered, often conflated with the former. This is unfortunate for several reasons. For one thing, the two intervals, while related, provide distinct types of information. The confidence interval defines the

range within which the population effect size lies (with a specified probability). In contradistinction, the prediction interval defines the range within which the effect size of future studies lies (with a specified probability). The latter is particularly relevant when the intent is to generalize the findings of the meta-analysis beyond the set of studies included in the review. A more detailed discussion of the prediction interval and the generalizability of meta-analytic summaries is provided in the discussion in Chapter 9.

CHAPTER 8. THE INTEGRATED META-MODEL

Step 6: Develop an Integrated Meta-Model

The primary purpose of the final integrated synthesis is to identify the different configurations of critical components—the causal recipes—driving the documented program effect. Towards this end, the findings from the implementation and effectiveness meta-summaries are integrated using Qualitative Comparative Analysis (QCA). As described above, the QCA involves six steps in the present synthesis:

1. Recode effectiveness studies according to the critical components identified in the implementation studies (component coding);
2. Develop a component/outcome matrix;
3. Use QCA software to create a ‘truth table’;
4. Minimize solution;
5. Resolve contradictory configurations;
6. Present final interpretation of solutions.

In the following, these steps are illustrated.

Step 1. Component coding

In the first step, the identified effectiveness studies are recoded according to the five critical components identified in the implementation meta-summary: low client/staff ratio, 24-hours crisis services, team approach, in-vivo service delivery, and time-unlimited services. Recall that these program components were identified as potential drivers of program effects in the implementation meta-summary described above. Towards recoding the effectiveness studies, the

five components were first operationalized into a four-value fuzzy set, with codes ranging from *no implementation* (0) to *low implementation* (.33), *high implementation* (.66), and *full implementation* (1). The coding scheme, including verbal anchors for each code, is shown in Table 8.1.

In addition to the component coding, the study outcomes (i.e., effect size estimates from the effectiveness meta-summary) were recoded according to three categories: small effect, moderate effect, and large effect. These benchmarks were based on the distribution of mean effect sizes and corresponding confidence intervals presented in Table 7.4 above. As noted by Schneider and Wagemann (2013), the outcome scores in QCA should always emerge from a qualitative judgment. For instance, and as evidenced in the present case, simply relying on the individual effect size estimates would fail to account for the information provided in the corresponding confidence intervals, which offers important information about the plausible range of the true effect size. As such, these ranges should be included in the outcome judgments for the purpose of QCA.

Table 8.1 Component coding scheme for ACT programs

Study	Full Implementation (1)	High Implementation (0.66)	Low Implementation (0.33)	No Implementation (0)
Low client/staff ratio	CSR below 10	CSR 11-20	CSR 21-30	CSR over 30
24-hour services	Crisis services 24 hours a day, seven days a week.	Crisis services 24 hours a day or seven days a week (but not both)	Crisis services neither 24 hours a day nor seven days a week	No crisis services
Team approach	Entire team responsible and provide all services for each client	Entire team responsible or all services provided by team (but not both)	Limited team responsibility and limited services provided by team	One team member responsible, most services provided outside team
In-vivo service delivery	All services delivered in-vivo	Most services delivered in-vivo	Some but not all services delivered in-vivo	No services delivered in-vivo
Time-unlimited services	Time-unlimited services for all clients	Time-unlimited services for most clients	Time-unlimited services for some clients	No time-unlimited services

Step 2. Develop component/outcome matrix

The results of the recoding are provided in Table 8.2. Each row represents an effectiveness study.

In the first five columns, the extent to which the ACT program in each study implemented a critical component is noted. The degree of implementation ranges from no implementation (0), through low (0.33), to high implementation (0.66) and finally full implementation (1). The sixth column provides the recoded outcome for each study. For instance, the ACT program in the study by Morse-St. Louis (1992) is characterized by full implementation of low client/staff ratio, team approach, in-vivo service delivery, and time-unlimited services as well as high implementation of 24-hour crisis services. The program outcome for this study is moderate.

Scanning across the seven studies and their recorded degree of implementation of the five core components reveals noticeable variations (though without any immediately detectable pattern).

Table 8.2 Component/outcome matrix for ACT programs

Study	Low client/staff ratio	24-hour services	Team approach	In-vivo service delivery	Time-unlimited services	Outcome
Morse-St Louis (1992)	1	0.66	1	1	1	Moderate
Lehman-Baltimore	0.66	0.66	0.66	0.66	0	Large
Morse-St. Louis (1997)	1	1	0.66	1	1	Large
Holter-New York	0.66	0.33	0	0.66	0	Moderate
Clarke-Portland	0.66	0.33	0.66	0.66	0.33	Small
Rosenheck-San Francisco	0.66	0.33	0.33	0.66	0	Small
Morse-St. Louis (2006)	1	0.33	0.66	0.66	0.33	Small

Steps 3-5. Use QCA software to create a ‘truth table’, minimize solutions, and resolve contradictions

In the third step of the QCA the values for each study (across the five components) are arranged in a truth table, displaying the logical configurations of causal conditions that elicit a positive outcome. In the present synthesis, FsQCA (a software developed by Ragin & Davey, 2014) was used to produce the truth table presented as Table 8.3 below. In this truth table, each row represents a specific configuration of critical components (i.e., ‘causal recipe’). The number of studies representing each component configuration is provided. To illustrate, the first row represents ACT programs with low client staff ratio and in-vivo service delivery—two of the studies reflect this combination. Finally, the consistency score expresses the “degree to which the

cases sharing a given condition or combination of conditions agree in displaying the outcome in question” (Ragin, 2006, p.2).

This part of QCA already reveals some interesting things about the ACT programs. First and foremost, a lot of possible program configurations are not covered by any of the seven programs in the present synthesis. This is not surprising considering the ratio of seven studies in relation to the 32 possible configurations of components (a truth table has by definition 2^k rows, with k being the number of causal components included in the model). Moreover, the truth table also allows insights into the typical programmatic composition of ACT programs. Especially noteworthy are the programs in row three and four that largely ‘agree’ in their implementation of the program components (with the exception of ‘time unlimited services’).

Finally, there are relatively high, but by no means perfect, consistency scores for several configurations, suggesting that studies within each combination generate a high positive effect. The only exception is the second row, where the consistency score at .67 is relatively low.

Table 8.3 Truth table for ACT programs

Low client/staff ratio	24-hour services	Team approach	In-vivo service delivery	Time-unlimited services	Number of studies	Consistency
1	0	0	1	0	2 (28%)	.85
1	0	1	1	0	2 (57%)	.67
1	1	1	1	1	2 (85%)	1
1	1	1	1	0	1 (100%)	1

In the subsequent step, the FsQCA software applies inferential logic (Boolean set algebra) to simplify the truth table into the causal recipes that are sufficient to produce a positive outcome. Stated differently, these simplified causal recipes can be viewed as the minimal combinations of causal conditions that constitute one of several possible paths to an outcome

(Ragin, 2006). The patterns in rows three and four may serve to illustrate this simplifying logic. Whereas the third row in the truth table indicates that implementation of all the five core components elicits a positive outcome, the fourth row indicates that implementation of the first four components elicits a positive outcome. Accordingly, the fifth component, time-unlimited services, is not a necessary condition for a positive outcome. A simplified—yet sufficient—causal recipe, across those two rows, is the program configuration of low client/staff ratio, 24-hour crisis services, a team approach, and in-vivo service delivery. This is in fact one of two proposed recipes in the present QCA.

The results of the truth table simplification are summarized in Table 8.4. As the table shows, the QCA revealed two causal recipes:

1. *Low client-staff ratio*24 hour crisis services*Team Approach*In vivo service delivery.* These are ACT programs with low client staff ratios and 24-hour crisis services combined with a team approach and in vivo service delivery.
2. *Low client-staff ratio*~24 hour crisis services*~Team Approach*In vivo service delivery*~time-unlimited services.* These are ACT programs combining low client-staff ratios and in-vivo service delivery but with no (or at least limited) 24-hour crisis services, team approach, and provision of time-unlimited services

These are two possible programmatic configurations that contribute to high housing tenure. In examining these, attention should first be awarded the coverage and consistency for each of the recipes. Following Ragin (2006), the coverage “assesses the degree to which a cause or causal combination ‘accounts for’ instances of an outcome” (p.2). In this way, the coverage reflects the empirical relevance of the recipe (the statistical equivalent would be R-squared). In a related vein, the unique coverage expresses the proportion of cases exclusively explained by the recipe.

Both ‘coverage’ and ‘unique coverage’ range from “0” (indicating no coverage) to “1” (indicating complete coverage). When there are several paths to the same outcome, the coverage of any given causal combination may be small. As noted above, the consistency of the recipes expresses the “degree to which the cases sharing a given condition or combination of conditions agree in displaying the outcome in question” (Ragin, 2006, p.2). Consistency ranges from “0” (indicating no consistency) to “1” (indicating perfect consistency).

Returning to Table 8.4 below, the first recipe displays noticeably higher degrees of coverage and consistency (.69 and 1, respectively), as compared with the second recipe (.46 and .85, respectively).² While there are no fixed standards for high consistency and coverage scores in QCA, a consistency score below .75 usually complicates any substantive conclusions and call for additional analyses of any program variations that may help explain the observed inconsistency (Ragin, 2006). These additional analyses may be in the form of reiterative rounds of program-specific analyses, recoding of studies on existing conditions, inclusion of new causal conditions, or any combination of these.

In the present context, the first recipe represents a more consistent and empirically relevant programmatic pathway to increased housing tenure. To better understand why the coverage and consistency scores differ across the two recipes, attention has to be awarded the studies that most closely align with each of the two recipes (in QCA terms these are studies with more than .5 membership in each recipe). For the first recipe, this would be studies with high implementation of low client/staff ratio, 24-hour crisis services, team approach, and in-vivo service delivery and with a strong positive outcome (i.e., coded with “.66” or “1” on each of

² Because fuzzy-set QCA allows for degrees of adherence to causal conditions, the consistency and coverage scores do not equate a distinct number of studies. Individual studies may hold degrees of membership to one or more recipes.

these). There are three studies with this programmatic configuration: Morse et al. (1992), Lehman et al. (1997), and Morse et al. (1997). These studies are perfectly consistent in the sense that they have high implementation on each of the four causal conditions and each display a moderate/strong housing tenure outcome.

In comparison, the second causal recipe is based on studies with high implementation of low client/staff ratio and in-vivo service delivery combined with low implementation of 24-hour crisis services, no team approach, and no time-unlimited services. The studies with greater than .5 membership in this causal recipe are Holter (1998) and Rosenheck et al. (2003). However, further examination of these studies reveals that the Rosenheck et al. (2003) study, while representing the causal recipe well, indicates a relatively small effect on housing tenure. As such, the consistency score for the recipe is reduced.

Table 8.4 Causal recipes for ACT programs (housing tenure)

	Coverage	Unique Coverage	Consistency
Client/staff ratio*24-hours crisis services*Team approach*In-vivo service delivery	.69	.38	1
Client/staff ratio*~24-hours crisis services*~Team approach*In-vivo service delivery*~Time-unlimited services	.46	.15	.85
Solution coverage: .84			
Solution consistency: .91			
Table note: Adapted from Legewie (2013)			

Step 6. Present and interpret the final set of causal recipes

Informed by the causal recipes identified in the QCA analysis, the final step of the integrated synthesis is to present and interpret the derived causal recipes. Developing this interpretation involves re-examining the specific studies contributing to the causal recipe(s) as well as reconsidering the mechanisms for the included critical components.

In the present context, only the first recipe is to be considered for further interpretation and presentation. Whereas the second causal recipe indicates a relatively high consistency (.85), the ACT programs in the two studies with primary membership in the recipe only resulted in small to moderate housing tenure outcomes. As such, the recipe is for the purpose of planning and designing effective ACT programs less compelling. In contradistinction, the first recipe, emphasizing high implementation of all but time-unlimited services, indicates a broader coverage and higher consistency. Moreover, the three studies with primary membership in the recipe indicate moderate to large housing tenure outcomes. As such, the first recipe demands further attention.

Towards this end, the underlying mechanism for each of the four components is worth re-examining. As indicated in Table 6.3 above, the mechanisms for the components revolve around care that is *intense* and *individualized* (i.e., low/client staff ratio), *direct* (i.e., team approach), *immediate* and *continuous* (i.e., 24-hour services, seven days of week), and provided in the *clients' community* (i.e., in-vivo service delivery). Taken collectively, the emphasis on care provision that is intensive, individualized, direct, and offered when and where clients need support, makes perfect sense in the context of homeless individuals with severe mental illness, a subpopulation that is notoriously difficult to engage and retain in continuous treatment. As such,

it is reasonable to consider these four program components critical ingredients of ACT programs for this specific subpopulation.

In summary, then, the present meta-modeling synthesis identifies one causal recipe, or programmatic combination of critical ingredients in ACT programs, that promotes housing tenure among homeless individuals with mental illness. More specifically, the recipe suggests that low client/staff ratio combined with provision of 24-hour crisis services, a team approach, and in-vivo service delivery, are sufficient critical ingredients for positive housing tenure. Taken collectively, these four program components promote service provision that is intense, individualized, direct, immediate and continuous, offered when and where the client is in need. As such, the program components are particularly salient in the context of care provision for homeless individuals with severe and persistent mental illness, which are often difficult to engage and retain in mental health treatment programs.

Methodological reflections

A couple of important points about the recoding in step one are called for. First and foremost, the coding is intentionally qualitative in the sense that each code represents a qualitative judgment by the researcher. A central aspect of this coding is the level of specificity called for by the coding scheme. In the present synthesis, a relatively simple four-value coding scheme was deployed. This is in the context of fuzzy-set QCA a relatively simple coding scheme. Whereas more fine-grained coding (for instance, using five- or seven-value coding schemes) would allow for more nuanced distinctions between ACT programs to be captured and analyzed, the program descriptions provided in published studies are unlikely to support the level of specificity in these types of coding schemes. In the present synthesis, the use of a four-value set was already made

difficult by the uneven program descriptions. Whereas some studies provided detailed information about the ACT program being studied, other studies offered less than a full paragraph to their program descriptions. Accordingly, the confidence with which the studies were coded varied from study to study. As such, pursuing a more fine-grained coding scheme would have resulted in air of precision unsupported by the available program information. In developing the coding scheme, the presence of relevant information (or lack thereof) tightly curbs the level of detail in the coding scheme as well as the confidence in the resultant coding.

A related limitation relates to the need for deep knowledge of the cases. The use of QCA represents a shift from a variable to a case-based treatment of the studies. A central part of this shift is the demand for more qualitative, in-depth information about the programs in the synthesis. However, this is not without difficulties in the context of applying QCA as part of a synthesis, where program information solely stems from published materials and where there is little opportunity to retrieve additional information. In the present synthesis, this difficulty emerged during the initial coding of the studies, as suggested in the preceding observation above. In order to adequately code the studies according to the five program components, program descriptions often had to be ‘pieced together’ across multiple publications on the same program or study. In its practical application, this involved identifying all published manuscripts on a given program and extracting program relevant information from these.

Another practical challenge relates to the highly iterative *modus operandi* of QCA, whereby the causal recipes emerge from the back-and-forth between within-case coding and cross-case comparisons. This central aspect of QCA runs against the grain of the linear approach in traditional syntheses and makes it difficult to sustain analytical transparency. To illustrate, the proposed causal recipes may display low levels of consistency, prompting the researcher to

revisit the cases contributing to the recipes, recoding these according to the included conditions, or even introducing new conditions to resolve any conflicts. In the present context, where conditions are included on the basis of the implementation meta-summary, any subsequent inclusion of additional conditions seems potentially unsound.

Finally, the causal recipe in the present synthesis was based on just three programs (or studies), two of which were produced by the same first author (Morse). Accordingly, concerns about the empirical grounding for the identified causal recipe could be raised. This is admittedly a challenge for which there likely is no clear-cut operational answer. Perhaps the best approach is to consider the identified causal recipe as a hypothesized causal recipe about how ACT programs for homeless individuals with severe mental illness works, treating the causal recipe as a subject for further empirical studies. As such, the development of causal recipes is less about theory confirmation and more about theory development.

CHAPTER 9. META-MODELING SOCIAL PROGRAMS—METHODOLOGICAL REFLECTIONS AND IMPLICATIONS

The overarching objective of the present dissertation is to advance the practical application of mixed methods syntheses. Towards this aim, the following questions guided the dissertation:

1. How do we systematically and transparently integrate findings from different types of studies?
2. What are the broader methodological implications of this type of integration for how we build knowledge across existing studies?

Returning to these important questions, the present chapter considers each of these in turn.

Towards addressing the first question, the chapter first reflects on the application of meta-modeling on ACT programs. More specifically, the methodological promises and pitfalls of meta-modeling are considered, with particular attention awarded to its systematic and transparent integration of findings from effectiveness and implementation studies.

Towards addressing the second question—pertaining to the broader methodological implications of meta-modeling for knowledge building—the second part of the chapter considers how the meta-modeling application on ACT may invite us rethink how and with what purpose we build knowledge across existing studies. More specifically, we may wish to reconsider some of the foundational assumptions underlying traditional approaches to systematic reviews and the kinds of knowledge to be built on the basis of these.

Finally, and informed by the first two sections, a set of propositions about future refinements and applications of meta-modeling as well as broader methodological topics to be pursued bring the dissertation to a close.

The Methodological Promises and Pitfalls of Meta-Modeling

A central component of the present dissertation is to illustrate the rationale and practical application of meta-modeling—an operational approach for mixed methods synthesis. In broad strokes, meta-modeling allows for a systematic and transparent integration of findings from program effectiveness and implementation studies. The approach, in its current stage of development, has several compelling features. First and foremost, the meta-modeling approach not only allows for the identification of the most salient critical ingredients in ACT programs (other synthesis approaches could potentially do this also), but it furthermore pushes for a more systematic integration of effectiveness and implementation studies. By combining findings from these different types of studies, meta-modeling allows for a deeper understanding of how and in what way programs work. This type of information serves well to inform the design and implementation of future versions of social programs as well as further research on these.

Second, meta-modeling offers a systematic and transparent treatment of the studies included in the synthesis. In line with traditional systematic reviews, meta-modeling follows a specified set of procedural steps that serves to ensure a systematic and transparent identification, extraction, coding, and analysis of the primary studies. The procedural steps of meta-modeling are made operational by way of concrete analytical strategies and data display techniques. In the final integration of findings, the use of Qualitative Comparative Analysis provides a visible chain of evidence, allowing for a verifiable and transparent integration. In this way, the meta-modeling approach speaks directly to commonly raised concerns about the lack of transparency and systematic procedures characterizing other approaches for mixed methods syntheses. This is from a methodological perspective an important contribution.

Third, and equally important, meta-modeling is both operational and pragmatic. The approach, while including a broader range of studies, from experimental to observational, from effectiveness to implementation studies, is made feasible by focusing on specific studies with specific types of information. This stands in marked contrast with other mixed methods synthesis approaches, such as realist synthesis, where a much broader range of quantitative, qualitative, or mixed method studies may be included. By focusing on specific types of studies, that is, effectiveness studies for estimating program effects and implementation studies for identifying critical components driving program effects, meta-modeling provides for a more focused and purposeful synthesis of different types of studies. Moreover, the synthesis stays true to the original intent and purpose of the primary studies, capitalizing on their comparative strengths.

No methodological approach is without its flaws. And meta-modeling is no exception. One practical limitation, and one which reaches beyond meta-modeling, concerns the difficulty of distinguishing between the implementation of the primary study and the reporting of the primary study. This concern emerges from the simple observation that many important aspects of studies are never reported. The lack of reporting on salient aspects of the program studied, which may be due to page limitations, journal guidelines, among other things, is particularly problematic in relation to the component coding of the effectiveness studies, as failure to code any of the candidate program components leaves the study ineligible for further analysis. The best strategy to counter limited program information in the published studies is to seek out additional information from program websites, other publications on the program, direct contact with authors of the studies, or even fieldwork on the program sites. However, all of these strategies are time-consuming, and some likely to be unfeasible within the time and resource constraints of most syntheses.

In a related vein, a second concern relates to the quality of the component coding for the QCA and in effect the robustness of the causal recipes identified. Would the recipes look different if the studies or the components of the studies had been coded with more confidence, or perhaps even differently? And how robust are the causal recipes? As Wagemann and Schneider remind us, “As a consequence of the explorative character of QCA results, they are always provisional, as they should always be followed by additional case studies and statistical or QCA analyses” (2007, p.18). From this perspective, the provisional nature of the illustrated QCA is in line with the present author’s general view on research syntheses: They tend to raise more questions than they answer. That being said, including some type of sensitivity analysis in future QCA-based syntheses, especially if applied towards more confirmatory purposes, would serve well to increase the confidence in the final conclusions.

Third and finally, the present application of meta-modeling did not include study design features, contextual conditions, or target group characteristics as part of the final integrated analysis. One potential strategy for doing so would be to include relevant design, target group, and contextual characteristics as conditions in the QCA, whereby the logical configurations of these conditions are examined along with program critical ingredients. This would allow for broader patterns to emerge across these different types of conditions. One challenge to this approach is that the increased number of conditions (features of the study design, context, and target group) relative to cases (number of studies) becomes exponentially disproportionate (a truth table has by definition 2^k rows, with k being the number of causal conditions included in the model). The end result would likely be that most of the rows (possible configurations of conditions) in the truth table would fail to be represented in any of the available cases (studies), a problem labeled “limited diversity” (Ragin, 1987, p. 106-113). Accordingly, pursuing this

strategy would require judicious selection of relevant characteristics. Moreover, the underlying assumption of this strategy is that relevant design, context, and target group characteristics are adequately described in the published studies. As illustrated above, the ability to code the studies in the present synthesis was severely restricted by limited and uneven reporting on program, study design, target group, and context characteristics. This challenge is not specific to meta-modeling, however, but common in relation to research syntheses more generally. The provision of more comprehensive program and context descriptions would surely enhance the analytical potential of research syntheses, meta-modeling included.

The Implications of Meta-Modeling for Knowledge Building

Broadening the scope of our methodological considerations, the illustrated application of meta-modeling also raises a number of questions about the purpose and role of knowledge building in evaluation, the implications of which are worth our consideration. Especially the development of the effectiveness meta-summary, for which meta-analysis served as the statistical workhorse, revealed a number of challenges, including a high amount of unexplained variation and relatedly a reliance on what might be considered an untenable assumption about the studies representing a random sample. Let us consider these in turn.

The heterogeneity challenge

The presence of unexplained variation is nothing new in meta-analysis. The very premise for and promise of meta-analysis is after all to summarize diverse findings from a broad pool of studies. As such, heterogeneity is inevitable (Higgins, 2002). In practical terms, the amount of heterogeneity may nudge the meta-analytic focus in three directions. Borenstein et al. (2009) neatly outlines these directions as follows:

1. If the effect size is consistent, then we will usually focus on the summary effect, and note that this effect is robust across the domain of studies included in the analysis.
2. If the effect sizes vary modestly, then we might still report the summary effect size but note that the true effect in any given study could be somewhat lower or higher than this value.
3. If the effect varies substantially from one study to the next, our attention will shift from the summary effect to the dispersion itself (p.7).

The meta-analysis illustrated above raised a number of issues in regards to at least two of these analytical directions. More specifically, it revealed a situation where the effect size variations left the meta-analytic focus somewhere between (1) and (3), without the ability to fully pursue either.

For one thing, the development of the effectiveness meta-summary revealed a high amount of heterogeneity, likely emerging from a combination of variations in program features, study designs, target group characteristics, and definitions of outcome measures, to name but a few. As a result, the combined effect size across all 15 effectiveness studies was rendered meaningless, as it was reflecting studies that were fundamentally incomparable.

Upon further examination of the effect sizes by subgroups (defined according to study design and program types), one subset of studies (seven experimental studies of regular ACT

programs) displayed similarly enough effect sizes to support a combined effect size estimate. Borrowing the formulation by Borenstein et al. (2009), the combined effect size was considered robust across the domain of the seven studies included in the subset. Whereas this combined effect size may be statistically sound, the reliance on a subset of seven studies (and the corresponding exclusion of the remaining eight studies from further analysis) is somewhat unsatisfying. This is at least in part because the domain of these seven studies seems overly restricted, anchoring the combined effect size to published experimental studies of ACT programs. With the potential aim of providing generalizable claims about ACT effectiveness, the reliance on such a narrow domain of studies appears less than ideal.

Given the large amounts of variation across the remaining eight studies, the only other analytical option is to statistically examine and adjust for the most salient variations affecting the program outcomes across the entire set of studies. This is equivalent to option (3) as formulated by Borenstein et al. (2009). However, pursuing this option also seems difficult in the present meta-analysis. For one thing, while it is possible to statistically explore and account for the effect size variations between studies, for example by way of meta-regression, these techniques are only possible for known and reported characteristics of the studies. Moreover, these techniques also demand what in many meta-analytic contexts is an unfeasibly high number of studies. As Thompson and Higgins (2002) advise, a methodologically sound application of meta-regression demands ten studies per covariate, assuming these covariates are equally distributed. Heeding their advice, conducting a sound meta-regression as part of the present synthesis, with salient variations across study designs, program features, target group characteristics, and outcome measures, among other things, would require at least 50 studies. And to add to the already tall order, these 50 studies would have to provide complete information on all covariates (Thompson

& Higgins, 2002). This is unfeasible in the present context and, as I would argue, likely to be difficult in the context of many a meta-analysis. As documented by Higgins, Thompson, Altman, and Deeks (2002), a recent review of 39 systematic reviews published under the Cochrane Collaboration (a coveted and authoritative repository of high quality systematic reviews) revealed only one systematic review with more than 15 primary studies. As such, the potential for meta-regression analyses is often limited. The demands of meta-regression seem unrealistic in the context of many, if not most, social programs.

The potential implications of the heterogeneity issue reach well beyond statistical concerns about the accuracy of effect size estimates. For one thing, and as indicated above, if a large amount of heterogeneity pervades the meta-analytic results, the estimation of a combined effect size makes little sense. However, identifying a smaller subset of homogenous studies on which to estimate a robust combined effect size runs the risk of restricting the domain of the effect size, which in turn limits the potential for generalizing the findings.

Another distinct implication, provided the studies support further statistical examination, is that the purpose of the meta-analytic summary shifts from determining an overall combined effect size to examining the dispersion of effect sizes (Borenstein's third direction above). From this perspective, the stated issue motivates a completely different conceptualization of the aim and promises of research synthesis. This shift is not necessarily a loss. The hard-earned clarity that may result from examining—as opposed to controlling for, or even worse, willfully ignoring—effect size variations provides a useful avenue for more nuanced insights about program effectiveness. Learning from variation becomes front and center. That being said, the shift still represents a departure from the traditional meta-analytic promise of producing a firm

and clear answer to the question: ‘what works?’ Perhaps by offering instead the only possible and authentic response: “it depends”.

In the end, heterogeneity is to be expected in the context of synthesizing effectiveness studies of social programs. As such, and unless future implementations of social programs are coupled with a surge of well-described experimental and quasi-experimental effectiveness studies, allowing in turn for statistical analyses of variations due to program- and study-level differences, alternative analytical avenues for examining effect size variations will have to be advanced. In this light, the future of mixed methods synthesis approaches, including the meta-modeling approach, looks bright. By offering a non-statistical alternative to the exploration of effect size variations, there could be an important role for these approaches in synthesizing social programs.

The random sampling challenge

Emerging in part from the forced analytical focus on the seven experimental studies in the present synthesis, another fundamental issue deserves our attention. This issue, which pertains to the generalizability of the present synthesis, cuts to the core of meta-analysis. To understand why that is, we first have to consider the three fundamental assumptions underlying meta-analysis.³

³ There are two main models for meta-analysis: the fixed-effect and the random-effects model. The fixed-effect model assumes that all observed effect sizes represent a single true effect size. Accordingly, any variation among the observed effect sizes stem from sampling error. In contradistinction, the random-effects model allows for multiple true effect sizes underlying the observed effect sizes, that is, the observed effect sizes may represent different true effects. Accordingly, any variation among the observed effect sizes may stem from true variation in effect sizes as well as sampling error. In the context of social programs, where program, target group, and other implementation variations across studies are to be expected, the assumption underlying the random-effects model is more appropriate. However, the random sampling concerns raised in the discussion are equally applicable to the fixed-effect model.

As formulated by Borenstein et al. (2009) these assumptions can be stated as follows:

1. There may be different true effect sizes underlying different studies;
2. These true effect sizes are normally distributed about some mean;
3. The observed effect sizes represent a random sample of these true effect sizes (p. 61).

The first two assumptions are easily motivated by the observed program and study-level variations across the ACT studies in the present synthesis. Whereas the ACT studies all pertain to ACT programs for homeless individuals with mental illness, the studies also display marked variations in program types, study features, target groups, among other factors. As noted by Borenstein et al., (2009),

When we decide to incorporate a group of studies in a meta-analysis, we assume that the studies have enough in common that it makes sense to synthesize the information, but there is generally no reason to assume they are identical in the sense that the true effect size is exactly the same in all the studies (p.61)

This view of studies as similar-enough-but-distinct can be illustrated as shown in Figure 9.1 below. As the figure shows, we might imagine three studies (each with an estimated effect size $\bar{x}_{1-3}$) to reflect three different true effect sizes (μ_{1-3}), corresponding to the programmatic differences across the three studies. The studies are not identical. The true effect sizes, however, are still assumed to reflect an overarching effect size, around which the true effect sizes are normally distributed. The three studies are still similar enough to support synthesis.

As evident in the present synthesis, especially the subgroup analyses, effect sizes may vary across studies in ways that suggests more than one true effect underlying the studies. Given the widespread diversity in both design and implementation of social programs, the first two assumptions seem acutely appropriate. The third assumption of a random sample appears equally

necessary for meta-analysis. This is because the random sampling assumption is the grounding for any type of statistical inference. Simply consider how any meta-analytical estimate of a combined effect size, any estimate of a corresponding error term and confidence interval, and any statistical significance tests or statistical generalization of same said combined effect size, rely entirely on the assumption of a random sample.

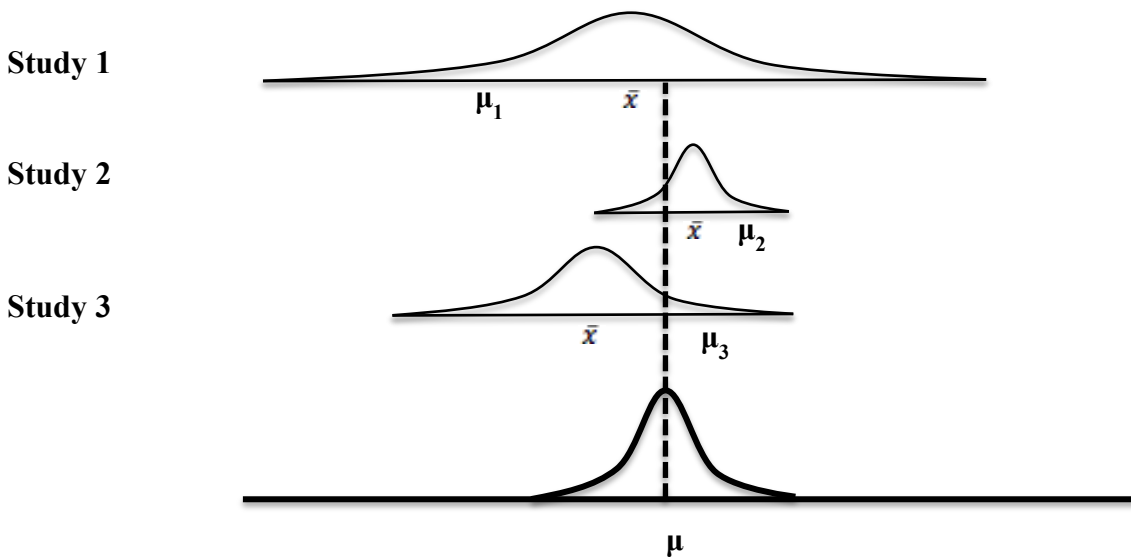


Figure 9.1 Random-effects model (adopted from Borenstein et al., 2009)

The issue with the random sampling assumption, however, is that it seems questionable, untenable even, in the context of social programs. To illustrate, the seven ACT programs contributing to the combined effect size estimate in the present synthesis, those studied by way of an experimental study, are likely to be systematically different from those ACT programs studied on the basis of quasi-experimental or non-experimental designs. The seven ACT programs were never randomly selected—from a sampling frame of all ACT programs—to receive an experimental study and subsequent publication. As such, they may be systematically

different from other programs. They might be better funded, more likely to be located in urban areas, more likely to be “signature programs,” target more homogenous populations, and so forth. To be sure, the seven experimental studies in the present review cannot be viewed as a random sample of all ACT programs. More likely, they represent ACT programs in published experimental studies and are as such better viewed as a smaller, purposively selected sample of the broader and diverse range of ACT programs.

One might argue that the seven studies still might be viewed as a random sample of all published experimental studies of ACT programs. However, this still assumes the studies were randomly sampled from this specified population, which they were not, and even if they were, their potential generalizability would be bordered within this narrow and arbitrary subset of the broader population of ACT programs. Moreover, the lack of randomization also pertains to the sampled participants within these studies. To illustrate, the homeless individuals with severe mental illness participating in the seven ACT programs were not randomly sampled from broader populations of homeless individuals with severe mental illness. The participants were referred by other agencies, actively recruited by the ACT programs, or in some situations self-enrolled. None of the ACT programs in the present set of studies randomly sampled their participants. As such, even the confidence intervals of the study-level effect sizes and in effect the generalizability of the latter seem questionable.

The lack of attention awarded the widespread (almost inevitable) violation of this fundamental assumption is troubling, somewhat surprising even, especially given the central role of generalizability in systematic reviews. As Matt and Cook (2009) remind us:

In our view, no rationale for meta-analysis is more important than its ability to identify the realm of application of a knowledge claim—that is, identifying whether the

association holds with specific populations of persons, settings, times and ways of varying the cause or measuring the effect; holds across different populations of people, settings, times, and ways of operationalizing a cause and effect; and can even be extrapolated to other populations of people, settings, times, causes, and effects than those studied to date. These are all generalization tasks that researchers face, perhaps no one more explicitly than the meta-analyst. (p.539).

From this perspective, and despite our general lack of attention and consideration, the implications of non-random sampling are realized and admittedly quite severe. For one thing, the violation of the assumption renders meta-analytic techniques for statistical inference, including confidence intervals and significance test, all but obsolete. Without random sampling, meta-analysis can no longer lean on the central limit theorem, or other principles of probability, that allows for statistical bridge building between a selective sample of studies and a broader population of interest. As such, the extent to which our sample-based effect sizes represent true effect sizes in a broader population is impossible to ascertain. There is simply no method by which to quantify and in effect assess the degree of uncertainty within the traditional meta-analytic toolbox. Consequently, our meta-analytic conclusions are tightly curbed to the specific set of individuals, programs, contexts, and settings, contained in our study sample. The validity of any claims reaching beyond the sampled studies is diminished beyond statistical repair.

In this way, the non-random sampling also holds implications for *how* and *with what purpose* we synthesize findings from different types of studies—the primary topic of the present dissertation. If our sample of studies is more purposive—than random—then our knowledge building becomes less generalizable (at least in the statistical sense of the word) and more situated within the boundaries of a specific set of studies. This does not mean that our syntheses

no longer support generalizable findings. What it does mean is that we will have to rely on alternative, non-statistical strategies for generalization. Matt and Cook (2009) propose several viable strategies, including: Establishing proximal similarity (i.e., applying effect sizes from the sample to other programs, settings, and target groups deemed similar to those of the sample), probing discriminant validity (i.e., examining how robust or sensitive the effect sizes are across variations in programs, settings, and target groups), and building on causal explanations (i.e., understanding the underlying processes mediating the program effects). These are strategies worth pursuing.

Taken collectively, these strategies also come with noteworthy implications for the type of knowledge we can build on the basis of syntheses. For one thing, the way in which we generalize findings will have to be more theory-based than statistically driven. Instead of relying on statistical generalizations from samples to populations, we will have to ground our generalization on a more fine-grained understanding of how social programs work. Recent work by Cartwright and Hardie (2012), may clarify what this entails. Following Cartwright and Hardie (2012), going from the program “works there” to how will it “work here” is grounded on two types of information: (1) Information about the causal principles (in the present context causal recipes) generating the program outcomes and (2) information about the most salient support factors for the program (perhaps best understood as contextual factors that will support program effectiveness). Advancing our knowledge building towards this type of theoretical generalization, then, requires that we specify the causal recipes and support factors that drive the desired outcomes, whereby generalizations of how the programs might work in other contexts can be made.

This shift in our foundation for generalization also implies a shift in the very purpose of knowledge building. This is because our focus no longer pertains exclusively to effect size estimates (i.e., generalizing outcomes) but instead centers on the underlying causal recipes generating these (i.e., generalizing critical program components). By better understanding the latter, we are able to at least theorize about how the program under review will function and perform in future settings, with future populations, and in future contexts. From this perspective, we will be able to transfer insights about how a program “worked there” to how it will “work here” (Cartwright & Hardie, 2012). From this perspective, meta-modeling, among other mixed methods syntheses, may have an important role to play in how we build knowledge from “worked there” to will it “work here.”

Moving Forward

The present dissertation is at its core about building knowledge across different types of studies. With the above considerations as a backdrop, the potential role and promise of meta-modeling is only made all the more compelling. To be sure, the challenges related to both statistical analyses of heterogeneity and statistical generalization might serve well to motivate further interest in meta-modeling. To illustrate, the emphasis on identifying the critical ingredients of social programs—especially the different configurations of these—serves well as an analytical foundation for better understanding how programs work, whereby theoretical generalizations extending beyond the initial set of studies can be made. This explicit focus on causal recipes is an important step towards the causal structures called for by Cartwright and Hardie (2012). Moreover, and by pursuing some of the suggested refinements proposed as part of the above

discussion of future applications of meta-modeling, the ability to pursue more theory-driven knowledge building on the basis of syntheses could be greatly enhanced.

Towards this aim, however, further refinements and applications are called for. For one thing, the meta-modeling approach needs to be implemented, tested, and refined in the context of other programs, other settings, at other times, and with other studies. The true test of any methodological development resides in its real world application. The meta-modeling approach is no exception. To earn a place among the burgeoning array of mixed methods synthesis approaches it has to show comparative methodological and practical advantages. Accordingly, future applications and modifications of the approach are encouraged.

Furthermore, future applications might refine and enhance meta-modeling in several important ways. In its current form, the meta-modeling approach stops short of explicating the underlying mechanisms generating the observed program outcomes. Whereas the present approach supports the identification of specific configurations of critical ingredients that elicit a positive outcome, it stops short of making explicit the underlying program mechanisms connecting these ingredients with specific outcomes. Accordingly, one analytical improvement could be the development of coding and analytical techniques to derive the causal mechanisms that serve to explain how and why the program, or perhaps even specific critical ingredients, work. By pushing the analysis in this direction, the synthesis could potentially foster an even deeper understanding of how and why programs work.

In line with the sharpened focus on mechanisms, future applications of meta-modeling could also place emphasis on a stronger integration of program theories in the development of causal explanations for the causal recipes identified. To illustrate, the systematic coding and analysis of program theories provided in published studies could potentially serve well to

formulate more fine-grained explanations about the underlying mechanisms of the program under study (or even specific critical ingredients). This type of theory-driven synthesis, using program theories to formulate and specify program mechanisms, might provide a better understanding of program effectiveness, which in turn could support theoretical generalizations reaching beyond the specific domain of studies included in the synthesis.

Finally, and in extension of the emphasis on mechanisms and program theories, future meta-modeling applications could further refine the analytical strategies relied on to empirically examine and test the causal structures underlying program effectiveness. The application of structural equation modeling, in particular, seems relevant in this regard. As proposed by Shadish (1996), statistically examining mediators (i.e., generative mechanisms) and moderators (e.g., gender or ethnicity) can change the interpretation of meta-analytic results, yielding important insights into program effects that are indirect or contingent (on one or more moderators). As such, structural equation modeling might serve to expand upon, refine, or even just corroborate the causal recipes identified in the Qualitative Comparative Analysis, and in this way support a more confident and confirmatory conclusion to the synthesis.

Moving beyond these methodological advancements of meta-modeling, the dissertation also motivates further work on the broader topic of knowledge building in evaluation. For one thing, the central role and feasibility of the random sampling assumption deserves further attention. If this foundational assumption is untenable in the context of meta-analyzing social programs, the potential implications of non-random study samples become topical. Also of interest are possible ways to remedy the non-random sampling. To illustrate, the potential role of non-parametric approaches for estimating individual and summarizing combined effect sizes is worth pursuing. As just one example, the application of traditional versus non-parametric meta-

analytic on the same set of studies might serve to illustrate the extent to which assuming random versus non-random sampling might influence the results and conclusions of a meta-analysis.

Another, and again broader, topic to be pursued cuts to the core of how and with what purpose we generalize findings from mixed methods syntheses. As suggested in the above discussion, the traditional notion of statistical generalizability is both in theory and practice unable to reach beyond the sample of studies included in the synthesis. Alternative strategies are called for. Work in this area would have to consider foundational distinctions in regards to the purpose of knowledge building (i.e., establishing versus explaining causal connections), the nature of generalization (i.e., statistical generalizability versus theoretical transferability of findings), and the subject matter of knowledge building (i.e., mechanisms versus outcomes). These topics, while by no means novel, have yet to be formalized into distinct and operational strategies for generalizing findings across evaluations. Specifying the operational procedures of a broader range of analytical strategies to support different types of generalizations are called for.

Finally, both of the above topics should be re-connected with the nature of social programming, that is, the way in which social programs are developed, disseminated, and, if deemed effective in the era of evidence, become labeled evidence-based practice. To illustrate, the evidence base for ACT programs appear to have developed across sequential generations of studies from early experimental studies serving to establish the efficacy of ACT; to cost-effectiveness studies establishing ACT as a viable care management model; to fidelity of implementation studies reflecting the growing dissemination of ACT programs; and finally to effectiveness studies of ACT programs covering a broader range of settings, populations, and contexts. Recognizing these distinct generations matters for how we build knowledge because each generation serves to structure and shape the scope and content of the evidence base. Our

ability to build knowledge is entirely defined by these generations of social program development. As such, the interplay between social program development and knowledge building (with all its potential limitations and opportunities) demands further consideration.

REFERENCES

*Manuscripts included in the synthesis

- Audini, B., Marks, I. M., Lawrence, R. E., Connolly, J., & Watts, V. (1994). Home-based versus out-patient/in-patient care for people with serious mental illness. Phase II of a controlled study. *British Journal of Psychiatry*, 165(2), 204-210.
- Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, 33, 5-18.
- Bond, G. R., Drake, R. E., Mueser, K. T., & Latimer, E. (2001). Assertive community treatment for people with severe mental illness: Critical ingredients and impact on patients. *Disease Management & Health Outcomes*, 9(3), 141-159.
- Bond, G. R., McGrew, J. H., & Fekete, D. M. (1995). Assertive outreach for frequent users of psychiatric hospitals: A meta-analysis. *The Journal of Mental Health Administration*, 22(1), 4-16.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. UK: John Wiley & Sons.
- Bronson, D. E., & Davis, T. S. (2012). *Finding and evaluating evidence. Systematic reviews and evidence-based practice*. New York, NY: Oxford University Press.
- Burns, B. J., & Santos, A. B. (1995). Assertive community treatment: An update of randomized trials. *Psychiatric Services*, 46(7), 669-675.
- Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: The Guilford Press.
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy—a practical guide to doing it better*. New York, NY: Oxford University Press.

- Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis, *Evaluation and the Health Professions*, 25(1), 12-37.
- Chalmers, T., & Lau, J. (1994). What is meta-analysis? *Emergency Care Research Institute*, 12, 1-5.
- Chandler, D., Meisel, J., McGowen, M., Mintz, J., & Madison, K. (1996). Client outcomes in two model capitated integrated service agencies. *Psychiatric Services*, 47(2), 175-180.
- *Clark, C., Guenther, C. C., & Mitchell, J. N. (2016). Case management models in permanent supported housing programs for people with complex behavioral issues who are homeless. *Journal of Dual Diagnosis*, 12(2), 185-192.
- *Clarke, G. N., Herinckx, H. A., Kinney, R. F., Paulson, R. I., Cutler, D. L., Lewis, K., & Oxman, E. (2000). Psychiatric hospitalizations, arrests, emergency room visits, and homelessness of clients with serious and persistent mental illness: findings from a randomized trial of two ACT programs vs. usual care. *Mental Health Services Research*, 2(3), 155-164.
- Coldwell, C. M., & Bender, W. S. (2007). The effectiveness of Assertive Community Treatment for homeless populations with severe mental illness: a meta-analysis. *American Journal of Psychiatry*, 164(3), 393-399.
- Cooper, H. (2010). *Research synthesis and meta-analysis—a step-by-step approach* (4th Edition). Thousand Oaks, CA: Sage Publications.
- Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd edition). New York, NY: Russell Sage Foundation.

- Cooper, H., Hedges, L. V., & Valentine J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd edition). New York, NY: Russell Sage Foundation.
- Dickersin, K., Min, Y. I., & Meinert, C. L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA*, 267, 374-378.
- Dixon, L. (2000). Assertive Community Treatment: Twenty-five year of gold. *Psychiatric Services*, 51(6), 759-765.
- Dixon, L., Friedman, N., & Lehman, A. (1993). Compliance of homeless mentally ill persons with assertive community treatment. *Hospital & Community Psychiatry*, 44(6), 581-583.
- Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research and Policy*, 10(1), 45-53.
- Dixon-Woods, M., Bonas, S., Booth, A., Jones, D., Miller, T., Sutton, A. J., Shaw, R. L., Smith, J. A., & Young, B. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research*, 6, 27-44.
- Dixon-Woods, M., Fitzpatrick, R., & Roberts, K. (2001). Including qualitative research in systematic reviews: Problems and opportunities. *Journal of Evaluation in Clinical Practice*, 7, 125-133.
- Drake, R. E., McHugo, G. J., Clark, R. E., Teague, G. B., Xie, H., Miles, K., & Ackerson, T. H. (1998). Assertive community treatment for patients with co-occurring severe mental illness and substance use disorder: a clinical trial. *American Journal of Orthopsychiatry*, 68(2), 201-215.

- *Drake, R. E., Yovetich, N. A., Bebout, R. R., Harris, M., & McHugo, G. J. (1997). Integrated treatment for dually diagnosed homeless adults. *The Journal of Nervous and Mental Disease, 185*(5), 298-305.
- Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337*(8746), 867-872.
- EPPI-Centre (2006). *EPPI-Centre methods for conducting systematic reviews*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
- Fekete, D. M., Bond, G. R., McDonel, E. C., Salyers, M., Chen, A., & Miller, L. (1998). Rural intensive case management: a controlled study. *Psychiatric Rehabilitation Journal, 21*, 4, 371- 379.
- Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education, 4*, 86-102.
- *Fletcher, T. D., Cunningham, J. L., Calsyn, R. J., Morse, G. A., & Klinkenberg, W. D. (2008). Evaluation of treatment programs for dual disorder individuals: Modeling longitudinal and mediation effects. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(4), 319-336.
- Glass, G. (1976). Primary, secondary, and meta-analysis of research. *The Educational Researcher, 10*, 3-8.
- Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage Publications.
- Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: Systematic review and recommendations. *The Milbank Quarterly, 82*(4), 581-629.

- Hak, T., Van Rhee, H. J., & Suurmond, R. (2016). How to interpret results of meta-analysis (version 1.0). Retrieved from: <http://www.erim.eur.nl/research-support/meta-essentials/downloads/>
- Harden, A., & Thomas, J. (2005). Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methodology*, 8, 257-271.
- Harden, A., & Thomas, J. (2010). Mixed methods and systematic reviews: examples and emerging issues. In A. Tashakkori & C. Teddlie (Eds.), *SAGE handbook of mixed methods in social and behavioral research* (2nd ed., 749-774). Thousand Oaks, CA: Sage Publications.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Herdelin, A. C., & Scott, D. L. (1999). Experimental studies of the program of Assertive Community Treatment (PACT). *Journal of Disability Policy Studies*, 10(1), 53-89.
- Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, Sussex: Wiley-Blackwell.
- Higgins J. P. T., Thompson, S. G., Altman, D. G., & Deeks J. J. (2002). Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Services Research and Practice*, 7, 51-61.
- *Holter, M. C. (1998). An intervention to reduce recurrent homelessness among severely mentally ill men: Benefit-Cost Analysis (unpublished doctoral dissertation). Columbia University, New York.
- Hoult, J., & Reynolds, I. (1984). Schizophrenia: a comparative trial of community oriented and hospital oriented psychiatric care. *Acta Psychiatrica Scandinavia*, 69(5), 359-372.

- Hunt, M. (1997). *How science takes stock: the story of meta-analysis*. New York, NY: Russell Sage Foundation.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis – cumulating research findings across studies*. Beverly Hills, CA: Sage Publications.
- Jackson, G. B. (1978). Methods for reviewing and integrating research in the social sciences. Final Report to the National Science Foundation for Grant No. DIS 76-20309. Washington, DC: Social Research Group, George Washington University.
- Jerrell, J. M., & Hu, T-W. (1989). Cost-effectiveness of intensive clinical and case management compared with an existing system of care. *Inquiry*, 26(2), 224-234.
- *Kenny, D. A., Calsyn, R. J., Morse, G. A., Klinkenberg, W. D., Winter, J. P., & Trusty, M. L. (2004). Evaluation of treatment programs for persons with severe mental illness. Moderator and mediator effects. *Evaluation Review*, 28(4), 294-324.
- *Korr, W. S., & Joseph, A. (1995). Housing the homeless mentally ill: Findings from Chicago. *Journal of Social Services Research*, 21, 53-68.
- Lafave, H. G., deSouza, H. R., & Gerber, G. J. (1996). Assertive Community Treatment of severe mental illness: a Canadian experience. *Psychiatric Services*, 47(7), 757-759.
- Latimer, E. A. (1999). Economic Impacts of Assertive Community Treatment: A review of the literature. *Canadian Journal of Psychiatry*, 44, 443-454.
- Legewie, N. (2013). An introduction to applied data analysis with qualitative comparative analysis (QCA). *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 14(3). Retrieved from: <http://www.qualitative-research.net/index.php/fqs/article/view/1961/3594>

- *Lehman, A. F., Dixon, L. B., DeForge, B. R., & Kernan, E. (1995). Assertive community treatment for homeless persons with schizophrenia and other chronic mental disorders. *Schizophrenia Research*, 15(1-2), 218-219.
- *Lehman, A. F., Dixon, L. B., Hock, J. S., DeForge, B. R., Kernan, E., & Frank, R. (1999). Cost effectiveness of assertive community treatment for homeless persons with severe mental illness. *British Journal of Psychiatry*, 174, 346-352.
- *Lehman, A. F., Dixon, L. B., Kernan, E., DeForge, B. R., & Postrado, L. T. (1997). A randomized trial of assertive community treatment for homeless persons with severe mental illness. *Archives of General Psychiatry*, 54, 1038-1043.
- Lemire, S. (under review). Meta-modeling Housing First: A mixed-methods synthesis approach. *Evaluation Review*.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University.
- Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among research studies. *Harvard Educational Review*, 41(4), 429-471.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
- Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.

- *Mares, A. S., & Rosenheck, R. A. (2011). A comparison of treatment outcomes among chronically homelessness adults receiving comprehensive housing and health care services versus usual local care. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(6), 459-475.
- Marshall, M., & Creed, F. (2000). Assertive community treatment—is it the future of community care in the UK? *International Review of Psychiatry*, 12(3), 191-196.
- Marshall, M., & Lockwood, A. (1998). Assertive community treatment for people with severe mental disorders (review). *The Cochrane Collaboration*.
- Marx, A. J., Test, M. A., & Stein, L. I. (1973). Extra-hospital management of severe mental illness. *Archives of General Psychiatry*, 29, 505-511.
- Mayne, J. (2012). Contribution analysis: coming of age? *Evaluation*, 18(3), 270-280.
- Matt, G. E., & Cook, T. D. (2009). Threats to the validity of research syntheses. In H. Cooper, L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.
- *McGrew, J. H., & Bond, G. R. (1995). Critical ingredients of Assertive Community Treatment: judgments of the experts. *The Journal of Mental Health Administration*, 22(2), 113-125.
- McGrew, J. H., Bond, G. R., Dietzen, L., McKasson, M., & Miller, L. D. (1995). A multisite study of client outcomes in assertive community treatment. *Psychiatric Services*, 46(7), 696-701.
- *McGrew, J. H., Bond, G. R., Dietzen, L., & Salyers, M. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology*, 62(4), 670-678.

- McHugo, G. J., Hargreaves, W., Drake, R. E., Clark, R. E., Xie, H., Bond, G. R., & Burns, B. J. (1998). Methodological issues in Assertive Community Treatment studies. *American Journal of Orthopsychiatry—Mental Health and Social Justice*, 68(2), 246-260.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis. A methods sourcebook* (third edition). Thousand Oaks, CA: Sage.
- Miller, N., & Pollock, V. (1994). Meta-analytic synthesis for theory development. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.
- *Monroe-DeVita, M., Teague, G. B., & Moser, L. L. (2011). The TMACT: A new tool for measuring fidelity to Assertive Community Treatment. *Journal of American Psychiatric Nurses Association*, 17(1), 17-29.
- *Morris, D. W., & Warnock, J. K. (2001). Effectiveness of a mobile outreach and crisis services unit in reducing psychiatric symptoms in a population of homeless persons with severe mental illness. *The Journal of the Oklahoma State Medical Association*, 94(8), 343-346.
- *Morse, G. A., Calsyn, R. J., Allen, G., Tempelhoff, B., & Smith, R. (1992). Experimental comparison of the effects of three treatment programs for homeless mentally ill people. *Hospital and Community Psychiatry*, 43(10), 1005-1010.
- *Morse, G. A., Calsyn, R. J., Klinkenberg, W. D., Cunningham, J., & Lemming, M. R. (2008). Integrated treatment for homeless clients with dual disorders: A quasi-experimental evaluation. *Journal of Dual Diagnosis*, 4(3), 219-237.
- *Morse, G. A., Calsyn, R. J., Klinkenberg, W. D., Helminiak, T. W., Wolff, N., Drake, R. E., Yonker, R. D., Lama, G., Lemming, M. R., & McCudden, S. (2006). Treating homeless

- clients with severe mental illness and substance use disorders: costs and outcomes.
Community Mental Health Journal, 42(4), 377-404.
- *Morse, G. A., Calsyn, R. J., Klinkenberg, W. D., Trusty, M. L., Gerber, F., Smith, R., Tempelhoff, B., & Ahmad, L. (1997). An experimental comparison of three types of case management for homeless mentally ill persons. *Psychiatric Services*, 48, 497-503.
- Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of research on case management. *Schizophrenia Bulletin*, 24(1), 37-74.
- Oakley, A. (2002). Social Science and evidence-based everything: The case of education. *Educational Review*, 54, 277-286.
- Olfson, M. (1990). Assertive community treatment: An evaluation of the experimental evidence. *Hospital and Community Psychiatry*, 41(6), 634-641.
- Oliver, S., Harden, A., Rees, R., Shepherd, J., Brunton, G., Garcia, J., & Oakley, A. (2005). An emerging framework for integrating different types of evidence in systematic reviews for public policy. *Evaluation*, 11, 428-466.
- Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (p. 3-11). New York, NY: Russell Sage Foundation.
- Patton, (2001). Evaluation, knowledge management, best practices, and high quality lessons learned, *American Journal of Evaluation*, 22(3), 329-336.
- Pawson, R. (2006). *Evidence-based policy: A realist perspective*. London: Sage.
- Pawson, R., & Boaz, A. (2004). Evidence-based policy, theory-based synthesis, user-led review. (ESRC Research Methods Programme Project). Retrieved from:
www.evidencenetwork.org/project3.asp

- Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2004). Realist synthesis: An introduction. (ESRC Research Methods Programme Project). Retrieved from: <http://www.ccsr.ac.uk/methods/publications/documents/RMPmethods2.pdf>
- Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3, 1243-1246.
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: a practical guide*. Malden, MA: Oxford Blackwell Publishing.
- Pillemer, D. (1984). Conceptual issues in research synthesis. *Journal of Special Education*, 18, 27-40.
- Pope, C., Mays, N., & Popay, J. (2007). *Synthesizing qualitative and quantitative evidence: A guide to methods*. Milton Keynes, UK: Open University Press.
- Quinlivan, R., Hough, R., Crowell, A., Beach, C., Hofstetter, R., & Kenworthy, K. (1995). Service utilization and costs of care for severely mentally ill clients in an intensive case management program. *Psychiatric Services*, 46(4), 365-371.
- Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley, CA: University of California Press.
- Ragin, C. C. (2006). Set relations in social research: Evaluating their consistency and coverage. *Political Analysis*, 14(3), 291-310.
- Ragin, C. C. (2014). *The comparative method—moving beyond qualitative and quantitative strategies*. Oakland, CA: University of California Press.
- Ragin, C. C., & Davey, S. (2014). fs/QCA Computer Programme, Version 3.0. Irvine, CA: University of California.

- Rihoux, B., & Ragin, C. C. (2009). *Configurational comparative methods—qualitative comparative analysis (QCA) and related techniques*. Applied Social Research Methods Series, V.51. Thousand Oaks, CA: Sage.
- Rosen, A., Mueser, K. T., & Teesson, M. (2007). Assertive community treatment—issues from scientific and clinical literature with implications for practice. *Journal of Rehabilitation Research and Development*, 44(6), 813-826.
- *Rosenheck, R. A., Kaspro, W., Frisman, L., & Liu-Mares, W. (2003). Cost-effectiveness of supported housing for persons with mental illness. *Archives of General Psychiatry*, 60(9), 940-951.
- Rosenheck, R. A., Neale, M., Leaf, P., Milstein, R., & Frisman, L. (1995). Multisite experimental cost study of intensive psychiatric community care. *Schizophrenia Bulletin*, 21(1), 129-140.
- Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, 85(1), 185-193.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-64.
- Rycroft-Malone, J., McCormack, B., Hutchinson, A. M., DeCorby, K., Bucknall, T., Kent, B., Schultz, A., Snelgrove-Clarke, E., Stetler, C. B., Titler, M., Walin, L., & Wilson, V. (2012). Realist synthesis: Illustrating the method for implementation research. *Implementation Science*, 7(33).
- Saini, M., & Shlonsky, A. (2012). *Systematic synthesis of qualitative research*. New York, NY: Oxford University Press.

- Salyers, M. P., McGuire, A. B., Rollins, A. L., Bond, G. R., Mueser, K. T., & Macy, V. (2010). Integrating assertive community treatment and illness management and recovery for consumers with severe mental illness. *Community Mental Health Journal*, 46(4), 319-329.
- Sandelowski, M., & Barroso, J. (2007). *Handbook for synthesizing qualitative research*. New York, NY: Springer Publishing Company, Inc.
- Sandelowski, M., Voils, C., & Barroso, J. (2006). Defining and designing mixed research synthesis studies. *Research in the Schools*, 13, 29-40.
- Santos, A. B., Deci, P. A., Lachance, K. R., Dias, J. K., Sloop, T. B., Hiers, T. G., Bevilacqua, J. J. (1993). Providing assertive community treatment for severely mentally ill patients in a rural area. *Hospital & Community Psychiatry*, 44(1), 34-39.
- Schaedle, R. W., McGrew, J. H., Bond, G. R., & Epstein, I. (2002). A comparison of experts' perspectives on assertive community treatment and intensive case management. *Psychiatric Services*, 53(2), 207-210.
- Schneider, C. Q., & Wagemann, C. (2013). *Set-theoretic methods for the social sciences—A guide to qualitative comparative analysis*. New York, NY: Cambridge University Press.
- Scott-Little, C., Hamann, M. S., & Jurs, S. G. (2002). Evaluations of after-school programs: A meta-evaluation of methodologies and narrative synthesis of findings. *American Journal of Evaluation*, 23(4), 387-419.
- Shadish, W. (1995). Author judgements about works they cite: Three studies from psychological journals. *Social Studies of Science*, 25, 477-498.
- Shadish, W. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, 1, 47-65.

- Shadish, W., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64*, 1290-1305.
- Sheldon, T. (1998). An evidence-based resource in the social sciences, report of a scoping study for the Economic and Social Research Council (ESRC).
- Smith, M. L., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.
- Solomon, P., & Draine, J. (1995). One-year outcomes of a randomized trial of case management with seriously mentally ill clients leaving jail. *Evaluation Review, 19*(3), 256-273.
- *Somers, J. M., Moniruzzaman, A., & Palepu, A. (2015). Changes in daily substance use among people experiencing homelessness and mental illness: 24-month outcomes following randomization to Housing First or usual care. *Addiction, 110*, 1605-1614.
- Stein, L. I., & Test, M. A. (1980). Alternative to mental hospital treatment: I. Conceptual model, treatment program, and clinical evaluation. *Archives of General Psychiatry, 37*, 392-397.
- Taube, C. A., Morlock, L., Burns, B. J., & Santos, A. B. (1990). New directions in research on Assertive Community Treatment. *Psychiatric Services, 41*(6), 642-647.
- *Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in Assertive Community Treatment: development and use of a measure. *American Journal of Orthopsychiatry, 68*(2), 216-232.
- *Teague, G. B., Drake, R. E., & Ackerson, T. H. (1995). Evaluating use of continuous treatment teams for persons with mental illness and substance abuse. *Psychiatric Services, 46*(7), 689-695.

- Test, M. A. (1992). Training in community living. In R. P. Liberman (Ed.) *Handbook of psychiatric rehabilitation*. Upper Saddle River, NJ: Prentice Hall.
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 1-10.
- Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., Brunton, G., & Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. *British Medical Journal*, 328, 1010-1012.
- Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Systematic Reviews*, 3. Retrieved from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4079172/>
- Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21, 1559-1573.
- Van der Knaap, L. M., Leeuw, F., Bogaerts, S., & Nijssen, L. T. J. (2008). Combining Campbell Standards and the realist evaluation approach—the best of both worlds? *American Journal of Evaluation*, 29(1), 48-57.
- Van Rhee, H. J., Suurmond, R., & Hak, T. (2015). User manual for Meta-Essentials: Workbooks for meta-analysis (version 1.0). Retrieved from: <http://www.erim.eur.nl/research-support/meta-essentials/downloads/>
- Wachter, K. W., & Straf, M. L. (1990). *The future of meta-analysis*. New York, NY: Russell Sage Foundation.
- Wagemann, C., & Schneider, C. Q. (2007). Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. Unpublished manuscript. Retrieved from: <http://www.compasss.org/wpseries/WagemannSchneider2007.pdf>

- Witheridge, T. F. (1991). The “active ingredients” of assertive outreach. *New Directions for Mental Health Services*, 52, 47-64.
- *Wolff, N., Helminiak, T. W., Morse, G. A., Calsyn, R. J., Klinkenberg, W. D., & Trusty, M. L. (1997). Cost-effectiveness evaluation of three approaches to case management for homeless mentally ill clients. *American Journal of Psychiatry*, 154(3), 341-348.
- Wong, G., Pawson, R., & Owen, L. (2011). Policy guidance on threats to legislative interventions in public health: a realist synthesis. *BMC Public Health*, 11(222).
- Wright-Berryman, J. L., McGuire, A. B., & Salyers, M. P. (2011). A review of consumer-provided services on Assertive Community Treatment and intensive case management teams: Implications for future research and practice. *Journal of Psychiatric Nurses Association*, 17(1), 37-44.
- *Young, M. S., Clark, C., Moore, K., & Barrett, B. (2009). Comparing two service delivery models for homeless individuals with complex behavioral health needs: Preliminary data from two SAMHSA treatment for homeless studies. *Journal of Dual Diagnosis*, 5(3-4), 287-304.
- Ziguras, S., & Stuart, G. (2000). A meta-analysis of the effectiveness of mental health case management over 20 years. *Psychiatric Services*, 51(11), 1410-1421.