STATISTICAL ISSUES IN A META-ANALYSIS OF STUDIES OF INTEGRATED TREATMENT PROGRAMS FOR WOMEN WITH SUBSTANCE USE PROBLEMS AND THEIR CHILDREN

MASTER OF SCIENCE (2008) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Statistical Issues in a Meta-analysis of Studies of Integrated Treatment Programs for Women with Substance Use Problems and Their Children

AUTHOR: Jennifer Liu, B.Math. (University of Waterloo)

SUPERVISOR: Dr. Lehana Thabane

NUMBER OF PAGES: viii, 38

Abstract

Meta-analysis is a statistical technique for combining findings from independent studies. A meta-analysis was performed to evaluate the effectiveness of integrated treatment programs for women with substance use issues and their children. Primary outcomes included substance use, maternal and child well-being, length of treatment, and parenting. A total of 9 randomized controlled trials (RCTs) and 84 observational studies were included in the final analysis. The *p*-value and capture re-capture method, were used to combine studies using different measures of treatment effect and evaluate the completeness of the literature search, respectively. Modified weights incorporating study quality were used to assess the impact of study quality on treatment effects. We also conducted a sensitivity analysis of correlation coefficients on combined estimates as a method for handling missing data.

Study quality adjusted weighting and traditional inverse variance weights provided different results for combined estimates of birth weight outcomes measured by standardized mean difference. The results from weighting by study quality provided a statistically significant result with a combined estimate of 0.2644 (95% CI: 0.0860, 0.4428), while the traditional method gave a non-significant combined estimate of 0.3032 with (95% CI: -0.0725, 0.6788). The sensitivity analysis of correlation coefficients (r) on combined estimates of maternal depression effects were similar, with confidence intervals that narrowed as r increased. Values of r = 0.2, 0.5, 0.65, 0.75, and 0.85 gave corresponding results (with 95% CI) of 0.67 (-0.10, 1.45), 0.67 (0.04, 1.3), 0.67 (0.12, 1.2), 0.67 (0.18, 1.15), and 0.66 (0.25, 1.07). Robustness of the sensitivity analysis for

study quality weighting and choice of correlation coefficient on combined estimates revealed benefits of integrated treatment programs for birth weight outcomes and maternal depression.

Evidence of benefit for at least some of the clients was apparent for parenting attitude measured by the Adult-Adolescent Parenting Inventory (AAPI). Results for each subscale of the AAPI were reported by timing of assessments (≤ 4, 5-8, ≥ 9 months). Combined p-values were 0.0006, <0.0001, <<0.0001 for Inappropriate Expectations, 0.1938, 0.1656, <<0.0001 for Lack of Empathy, 0.0007, <0.0001, <<0.0001 for Corporal Punishment, 0.0352, 0.0002, <<0.0001 for Role Reversal, and 0.5178 (5-8 months) for Power Independence. There was insufficient evidence for concluding a significant effect of treatment on neonatal behavioural assessments measured by Apgar scores. Combined p-values of 0.6980 and 0.3294 were obtained for the 1-minute and 5-minute Apgar, respectively.

The number of missing articles estimated by the capture recapture method was 8 (95% CI: 2, 24), which suggests a 90% capture rate of all relevant world literature. This result indicates that a sufficient amount of studies were retrieved to avoid bias in the results of the meta-analysis.

Conclusions regarding the effectiveness of integrated treatment programs were limited by poor quality evidence from individual studies. We suggest the use of statistical methods such as the p-value, capture re-capture, study quality weighting, and sensitivity analysis of correlation coefficients to handle missing data to address meta-analytic research questions and direct higher quality research in the future.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Lehana Thabane, for the opportunity to work on this project and for his guidance and support throughout my Masters experience.

I am extremely grateful to the entire research group for embracing my role in this project. In particular, I would like to thank Dr. Alison Niccols, Dr. Karen Milligan, and Ainsley Smith, from whom I have learned a tremendous amount and thoroughly enjoyed working with. I am thankful to my committee members, Dr. Noori Akhtar-Danesh and Dr. Román Viveros-Aguilera, for their valuable advice. I would also like to thank Dr. Gary Foster and Emmy Cheng for our stimulating discussions.

Finally, I would like to thank my family and friends for their support and encouragement in completing my Masters. Thank you to Bonita So and Steve Ling for taking the time and interest to proofread my thesis. Special thanks to Chenglin Ye, my fellow classmate, for his continuous support and advice as a colleague and friend.

Table of Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Effects of Substance Use for Women and Their Children

Women who use substances during pregnancy and after childbirth are at greater risk for experiencing health related issues, as are their children. Substance use during pregnancy increases a child's risk for prematurity, low birth weight, impaired physical growth and development, behavioural problems, learning disabilities and substance use [1-2]. Continued substance use after childbirth may affect a mother's parenting capabilities and her ability to provide a stable, nurturing environment for her children [3]. Substance-using women may also experience health problems related to mental health, relationships, physical or sexual abuse [4-5], poor nutrition and deficits in social support [6-7].

1.1.2 Treatment for Substance-using Women and Their Children

It has been suggested by researchers, clinicians and policy makers that treatment programs for substance use should be integrated and need to address women's as well as children's needs [8]. Such integrated programs would address women's physical, social and mental health, and at the same time meet children's needs through parenting and childcare services. An example of an integrated program is graphically depicted in Figure

1.1, Appendix A. The recognition for this need has led to the development of numerous integrated treatment programs which are primarily offered in residential or outpatient settings.

1.1.3 Outcome Studies and Reviews

Many studies have investigated the effectiveness of integrated treatment programs and suggested positive maternal and child outcomes, particularly decreased substance use and improvement in mental health, nutrition, employment, parenting, birth outcomes and child development [9-10]. However, these studies are faced with methodological limitations such as small sample size [9] and varying study quality that have made results hard to interpret. Systematic reviews in treatment research have focused on outcomes related to women's substance use [11-12], but have neglected to address maternal well-being, child well-being and parenting outcomes. The reviews have failed to include studies evaluating integrated treatment programs using a wide spectrum of maternal and child outcomes [9,13-15].

1.2 Objective

This thesis is part of an extensive systematic review and meta-analysis designed to determine the effects of integrated treatment programs on substance use, maternal well-being, child well-being, parenting and length of treatment, the impact of treatment on outcomes for women and children moderated by program characteristics, client characteristics, and study quality variables, and lastly, to what extent feasibility factors were described.

The statistical objectives of this thesis are to (i) employ an appropriate method for combining studies using different measures of treatment effect, (ii) determine the completeness of the literature search, (iii) assess the impact of study quality on effects of integrated treatment programs and (iv) conduct a sensitivity analysis of correlation coefficients on combined estimates of treatment effect.

1.3 Scope of this Thesis Report

In the following chapters, I will discuss the methods for conducting the systematic review and meta-analysis, statistical methods adopted for analysis and the results. These chapters will be followed by a discussion of the interpretation of the results and related issues, leading up to closing remarks.

In Chapter 2, I discuss methods for searching and evaluating the literature, more specifically, sources of the literature search, search strategy and study selection. Primary outcomes are also described.

Chapter 3 provides descriptions of the statistical methods used to address the objectives of this thesis, including combining p-values, the capture re-capture method, non-traditional weighting incorporating study quality and dealing with missing data through a sensitivity analysis of correlation coefficients on combined effect size estimates.

Chapter 4 presents the results from the capture re-capture method for assessing completeness of the literature search and combining p-values and weighting by study quality for selected outcomes. Results from the sensitivity analysis of incorporating study

quality into weights and the choice of correlation coefficients on combined estimates of treatment effect are also presented.

Lastly, a discussion on the key findings is found in Chapter 5. Interpretations, comparisons to similar works, limitations and implications for clinical practice and future research are discussed in this chapter. Concluding remarks are provided in Chapter 6.

Chapter 2

Methods of Literature Review

2.1 Sources of Literature Search

An extensive literature search was conducted using databases including MedLine, Pubmed, Embase, PsycINFO, Dissertations Abstract International, CINAHL, Sociological Abstracts and Web of Science, applying no limitations to the timeframe. Relevant journals and reference sections of retrieved articles were hand searched to find additional studies. The search period for relevant journals was 1990 to 2007. Since the adoption of integrated programs is more recent, it is unlikely we would retrieve relevant articles prior to 1990. Programs that may have relevant unpublished data were contacted to attempt to retrieve studies that have not been published in peer-reviewed journals. Additional information such as other variables not reported or detailed information regarding the treatment program may be missing and attempts to retrieve this information were made through contacting researchers and programs. Lastly, integrated programs identified through national treatment registries were contacted and proceedings and programs of relevant conferences were hand-searched.

2.2 Search Strategy

The literature search was conducted using Medical Subject Headings (MeSH), which is the National Library of Medicine's controlled vocabulary designed for indexing and searching the MEDLINE/PubMED databases, enabling retrieval systems to provide

subject searching of data. The following are MeSH terms used in this search including (and related to) *substance use* (*i.e. substance abuse*, *addiction* or *alcoholism*), *intervention* (*i.e. treatment*, *therapeutic*, *rehabilitation* or *program*), *women* (*mothers*), *pregnancy* (*pregnant*), *children* (*infants*), *mental health*, *prenatal* and *parenting*. When exact search terms were not available, their MeSH equivalents were used.

2.3 Study Selection

To be included in the meta-analysis, a study must be reported in English and meet criteria regarding the study participants, treatment program, study design and outcome variables. Study participants must be women who are pregnant or parenting, and have substance use problems at baseline. The treatment program must provide at least one service directly addressing substance use and at least one service related to children 0-16, including children who are unborn. A study must employ a randomized controlled trial, quasi-experimental, cohort or cross-sectional design, and quantitative results must also be provided for at least one of the primary outcomes (substance use, maternal well-being, child well-being, parenting or length of treatment).

A study was excluded from the review and meta-analysis if the treatment program included treatment of males or women who were not pregnant or parenting, or was a smoking cessation program. Case studies and qualitative studies not including quantitative data were also not eligible.

2.4 Primary Outcomes and Study Variables

The primary outcomes of the meta-analysis were substance use, maternal well-being, child well-being, parenting and length of treatment. For the purposes of this thesis we will use selected outcomes including the Adult-Adolescent Parenting Inventory (AAPI), Apgar scores, birth weight and maternal depression. More details on the AAPI and Apgar scores are available in Section 3.2.

Chapter 3

Statistical Methods

3.1 Overview

In this section, I will describe the methods adopted for handling statistical issues that arose in this meta-analysis. I also discuss a sensitivity analysis of correlation coefficients on combined estimates of treatment effect as a method to handle missing data. The five primary outcomes (substance use, maternal well-being, child well-being, parenting and length of treatment) were assessed using a variety of outcome measures. Estimates of treatment effect were combined over specific outcome measures and selected outcome measures that illustrated statistical issues of interest will be used to demonstrate appropriate methods for handling them.

A typical meta-analysis combines study estimates of treatment effects to achieve one overall estimate. In some cases, it may not be feasible or appropriate to combine effect sizes in this way, particularly if study outcomes are not reported on the same scale, if studies do not report effect-sizes but report p-values and when study designs or treatment levels vary too much to be combined [16]. In these cases, methods for combining p-values from individual studies may be most suitable for summarizing the data.

The main goal of a systematic review is to capture and analyze all literature available on a particular topic in order to answer the research question. We will use the capture re-capture method to estimate the total number of articles that could not be

identified from the available sources. Although the capture re-capture idea was pioneered in ecology, recent applications include its use in estimating the horizon of all nephrology journals [17], as a method for assessing publication bias in systematic reviews [18] and as a stopping rule when searching in systematic reviews [19].

The variability in study design and quality in this meta-analysis was high, and hence it was of interest to assess the impact of study quality on effects of integrated treatment programs. Study quality in this analysis will be incorporated through weighting, where individual quality scores will be used (instead of pure measures of precision) to determine the change in overall effect of the intervention [20]. We will illustrate this method for birth weight outcomes, under the assumption of a random effects meta-analysis model.

The 9 RCTs and 84 observational studies were analyzed by group (RCTs and quasi-experimental), cohort and cross-sectional designs. The timing of assessments of outcome measures also varied greatly between studies, and some studies could not be combined because length of treatment is associated with outcomes [11]. Timing of assessments were broken down into ≤ 4, 5-8 and ≥ 9 month intervals, where analyses were conducted within each time frame. SAS 9.1 software was used to conduct all analyses.

3.2 Combining *p*-values

Under the null hypothesis, a *p*-value from a continuously distributed test statistic is uniformly distributed from zero to one. Thus *p*-values from different tests are identically distributed under the null hypothesis and may be combined. The null and alternative hypothesis for combining *p*-values is,

$$H_0: \theta_i=0 \text{ versus } H_a: \theta_i\neq 0 \qquad \text{for } i=1,\ldots,k$$

where θ_i represents the treatment effect in the i^{th} study. In group designs, θ_i represents the treatment effect between the experimental and control group, and in cohort studies it represents the treatment effect between two time points.

There are several methods proposed for combining *p*-values, some based on the uniform distribution and others based on statistical theory for random variables [21]. The Logit method (based on the latter) for combining *p*-values proposed by George in 1977 [22], uses the test statistic:

$$-\sum_{i=1}^{k} log(p_i/1-p_i)\ [k\,\pi^2\,(5k+2)/3(5k+4)]^{-1/2} \qquad (1)$$

Since $log(p/(1-p))$ has a logistic distribution under H_0, and the distribution of sum of logits is similar to the *t* distribution, George proposed to approximate the distribution of (1) with the *t* distribution with $5k+4$ degrees of freedom [22].

I illustrate the *p*-value method for combining Apgar scores for group designs and AAPI subscale scores for cohort studies. The Apgar score is designed to evaluate a newborn's physical condition after delivery [23], and is recorded 1 minute and 5 minutes after birth with scores ranging from 0 to 10, higher scores indicating better outcomes. The Adult-Adolescent Parenting Inventory (AAPI) is a 32-item questionnaire designed to assess parenting attitudes and beliefs associated with child abuse and neglect [24].

The AAPI subscale scores were reported in the form of means, effect sizes, *p*-values and *t*-test values, while Apgar scores were reported as means and proportions. Each measure of treatment effect was converted to a *p*-value, by first converting the effect size to a test statistic and then conducting a two-sided test.

3.3 Capture Re-capture Method

The capture re-capture idea involves selecting k independent samples of size n_i, $i=1,\ldots,k$, from a population of size N (unknown), where the goal is to estimate N (horizon) [17]. After each sample is selected, they are marked 'captured' and returned to the population to be sampled again later. By marking them, we are able to identify in subsequent samples if an item has previously been captured, and this forms the basis of estimating the horizon.

The databases we searched included MedLine, PubMED, Embase, PsycINFO, Dissertations Abstract International, CINAHL, Sociological Abstracts and Web of Science, capturing a total of 75 relevant articles. Given that we searched a large number of sources, the feasibility and interpretability of the modeling step would be challenged since the number of interaction terms would increase substantially as the number of databases increases. Observing the pattern of studies found in each database, we found that Dissertations Abstract International and Sociological Abstracts did not make substantial contributions to the literature search, capturing only one and three articles respectively – all of which were captured by other databases. CINAHL captured 15 articles, where each article captured was located in at least one other database. The MedLine and PubMED searches were closely related, where all but one article found in MedLine were found in PubMED. For these reasons, we did not expect the MedLine, Dissertations Abstract International, Sociological Abstracts and CINAHL databases to contribute much value to the modeling step and were therefore excluded.

The aim was to estimate the horizon of the number of articles, where each database is thought to retrieve an independent sample. When all k databases are searched,

it yields an incomplete contingency table with 2^k-1 known cells and one missing cell (see Appendix B), where the missing cell indicates the unknown number of articles not identified in any of the databases. Fitting an unsaturated log-linear model to this incomplete table provides an estimated expected value for the missing cell, as well as associated confidence limits [25]. The sum of the estimated value of the missing cell plus values of the known cells gives an estimate of the horizon.

To obtain the best model for estimating the missing cell, we start with a simple model containing four main effects, where each of A, B, C and D denote PubMED, Web of Science, Embase and PsycINFO, respectively, and each factor with corresponding levels i, j, k and l.

$$log\ \mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D}$$

where

μ_{ijkl} = expected cell frequency of the cases for cell $ijkl$ in the contingency table
u = the overall mean of the natural log of the expected frequencies
u_i^{A}, u_j^{B}, u_k^{C}, u_l^{D} = main effect of A, B, C, D
u_{ij}^{AB}, u_{ik}^{AC}, u_{il}^{AD}, u_{jk}^{BC}, u_{jl}^{BD}, u_{kl}^{CD} = two-way interaction effects
u_{ijk}^{ABC}, u_{ijl}^{ABD}, u_{ikl}^{ACD}, u_{jkl}^{BCD} = three-way interaction effects.

Log-linear models have a hierarchal structure, meaning that if higher order terms are included in the model, then all related lower order terms are also included [26]. One approach is to build the model up by adding significant interaction terms. Hypothesis tests are conducted to determine whether an interaction term is significant by using the difference in residual deviances (D) of the model corresponding to the alternative hypothesis and the model corresponding to the null hypothesis [26]. Thus the change in deviance (ΔD) has a χ^2 distribution with degrees of freedom equal to the difference in the

degrees of freedom of the two models. The modeling portion was conducted using a previously written macro for estimating the horizon in a capture re-capture problem [27] in SAS 9.1.

3.4 Classical Meta-analysis with Random Effects Model

Under the assumption of a random effects meta-analysis model, the treatment effect estimator $\hat{\theta}_i$ from each individual study are thought to be independent samples from a normal distribution $N(\theta, \tau^2)$ [28]. Thus, the total variability for study i is given by

$$v_i^* = w_i^{-1} + \hat{\tau}^2,$$

where w_i^{-1} is the within-study variance, and $\hat{\tau}^2$ is the estimated between-study variance [28]. The estimates of treatment effect from each study are calculated differently for group (RCTs, quasi-experimental) and non-group (cohort, cross-sectional) designs. Within each study design, estimates of treatment effect are calculated according to the type of data reported for a particular outcome measure. A list of effect sizes for the selected outcomes is available in Appendix B.

If we let $w_i^* = (w_i^{-1} + \tau^2)^{-1}$, then the overall estimate of treatment effect is calculated by weighting each individual treatment effect by w_i^*. The random effects model was chosen because if the between-study variance is small, the overall estimate of treatment effect and associated standard errors and confidence intervals will be similar to the ones obtained from the fixed effect model. The results are reported as pooled estimates with corresponding 95% confidence interval.

3.5 Weighting by Study Quality

The Newcastle-Ottawa Scale (NOS) [29] and the Jadad Scale [30] were chosen to assess the quality of observational studies and RCTs respectively. The NOS was designed to assess the quality of non-randomized studies, including separate scales for cohort and case-control studies. The NOS is on a scale from 0 to 9 and covers three main topics: the selection of study groups, comparability of groups and ascertainment of outcome of interest (cohort) and exposure (case-control) [29]. The content validity and inter-rater reliability for the NOS have been established, but further evaluation is still being conducted [29]. The Jadad Scale was designed to assess the quality of RCTs and is widely used in the medical literature for this purpose [31]. The Jadad is a scale from 0 to 5 and is based on three criteria: if the study is described as randomized, double-blinded, and if a description is provided for withdrawal and dropouts [30].

The new weights,

$$w_i' = (SQ_i)(w_i^*)$$

were used to weight individual studies, where SQ_i is the quality score and w_i^* is the original weight given in the random effects analysis from each study [20]. Since quality scores alone are not pure measures of precision, using them as weights in a meta-analysis lacks statistical and empirical justification [20]. Multiplying the quality score by the inverse variance allows us to incorporate both measures of precision and study quality into the weight. We illustrate the incorporation of study quality into weighting for birth weight outcomes in quasi-experimental studies.

3.6 Sensitivity Analysis of Correlation Coefficients in Estimating Study Level Variances

A large portion of the studies in this meta-analysis were observational with cohort design. A problem among cohort studies was that none of them reported the correlation coefficient r needed to estimate the inverse variance w^{-1} of the standardized mean gain d_{smg},

$$w^{-1} = 2n/[4(1-r) + d_{smg}^2],$$

where n represents the number of studies [32]. Thus we could not combine effect sizes since we had complete missingness for their associated inverse variances. In an attempt to handle such missing data, we conducted a sensitivity analysis of r to assess the impact of a range of values on combined estimate of treatment effect under the random effects meta-analysis model. Since we expected the correlation between outcomes between two time points to be high, we chose values of r between 0.5 and 0.9. In the case that correlation between two time points was not significant, we also chose a low value for r of 0.2.

Chapter 4

Results

4.1 Results of Combining *p*-values

4.1.1 AAPI Outcome

The AAPI is composed of six subscales, and cohort studies in this meta-analysis reported on Total Score, Inappropriate Expectations, Lack of Empathy, Corporal Punishment, Role Reversal and Power Independence. Because outcome measures were combined in categories according to timing of assessment, studies that did not report on timing of assessment were included in all time interval categories.

Only one study used Total Score as an outcome measure, and could not be combined with any of the other studies. The results of combining *p*-values for studies reporting Inappropriate Expectations, Lack of Empathy, Corporal Punishment, Role Reversal and Power Independence at ≤ 4, 5-8 and ≥ 9 months are summarized in the following table.

	Timing of Assessment (*number of studies*)		
Outcome	≤ 4 months	5-8 months	≥ 9 months
Inappropriate expectations	0.0006 (2)	<0.0001 (7)	<<0.0001 (5)
Lack of empathy	0.1938 (3)	0.1656 (8)	<<0.0001 (6)
Corporal punishment	0.0007 (3)	<0.0001 (8)	<<0.0001 (6)
Role reversal	0.0352 (3)	0.0002 (8)	<<0.0001 (6)
Power independence	-	0.5178 (2)	-

Table 4.1: Overall p-value estimate based on combining p-values for AAPI

AAPI measures for Inappropriate Expectations, Corporal Punishment and Role Reversal had significant combined p-values over all time intervals. Lack of Empathy had a significant combined p-value in only one time interval ($\geq$ 9 months). These combined significance tests reject the overall null hypothesis at the 0.05 level, indicating at least some of the population had AAPI measures at follow-up that deviated from the initial assessment. By examining our data, this indicates a positive change as a result of treatment. Power Independence at 5-8 months and lack of empathy at ≤ 4 and 5-8 months did not result in significant combined p-values, so there is no sufficient evidence to reject the null hypothesis that all clients experienced no effects at follow-up from the initial assessment. For the four subscales of the AAPI (Inappropriate Expectations, Lack of Empathy, Corporal Punishment and Role Reversal) that were measured over all three time intervals, the general trend showed that combined p-values decreased as the timing of assessment increased. This trend may be apparent due to the fact that length of treatment in previous research has been associated with better outcomes, providing stronger evidence to reject the null hypothesis.

4.1.2 Apgar Score

Apgar scores for group-design studies were combined by 1-minute and 5-minute assessment times. The 1-minute score was reported by 5 studies and the 5-minute score was reported by 4 studies. The combined p-values for both scores were not significant for either score, as the p-value for the 1-minute score was 0.6980 and 5-minute score was 0.3294. These combined significance tests provide no evidence to reject the overall null hypothesis at the 0.05 level, that all studies experienced the same effect in the treatment

and control groups. Non-significance of the test may have resulted because studies did not control for time of entry into the program, so women may have joined the study late in their pregnancy and not have been able to benefit from the full scope of the treatment program.

4.2 Results of Estimating the Horizon of Literature Search

The results of the modeling step give a final best model,

$$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{ik}^{AC} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD} + u_{ikl}^{ACD} \quad (2)$$

Interaction terms were added in the following sequence in Table 4.2, while all models considered are outlined in Appendix B.

	Model	Deviance	*p-value*
1	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D}$	47.01	
2	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{kl}^{CD}$	35.53	<0.0008
3	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{ik}^{BC} + u_{jk}^{CD}$	25.11	<0.0013
4	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD}$	16.94	<0.0043
5	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{ik}^{AC} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD}$	12.67	<0.0388
6	$log\,\mu_{ijkl} = u + u_i^{A} + u_j^{B} + u_k^{C} + u_l^{D} + u_{ik}^{AC} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD} + u_{ikl}^{ACD}$	7.70	<0.0258

Table 4.2: Results of the modeling step for estimating the horizon

The final model (2) estimated the missing number of studies, with associated 95% confidence interval, to be 8 (2, 24). Thus our horizon estimate is $N = 83$ with 95% confidence interval (77, 99), where the known number of studies covers 90% of the

horizon. Since we assume a closed population, in which there is no change to the size of the population, obtaining 90% of the horizon seems like a reasonably high capture rate, although there is no formal way to decide on a cutoff value. Investigators must decide whether enough articles have been obtained to avoid bias in the overall results of the meta-analysis, or whether additional searches for relevant studies should be conducted.

Hand-searching of relevant journals, reference sections of previously retrieved articles and conference proceedings were conducted in addition to searching databases. These searches re-captured several articles, but since we did not record the search history for hand-searching, we could not include it in the capture re-capture estimation of the horizon. Additional attempts to obtain 'fugitive' literature through contacting relevant programs and researchers results in 19 additional articles found. These included special reports that were conducted primarily for internal program evaluations that would not be published or indexed in relevant databases. This brought the total number of articles retrieved to 94, before estimation of the horizon. Since an extensive search was conducted to retrieve all relevant data, published and unpublished, we do not suspect our conclusions to be biased as a result of restricting to published studies.

4.3 Results of Weighting Studies using Study Quality

There were 7 (non-RCT) studies in this meta-analysis in which birth weight outcomes were reported, with study quality scores on the NOS ranging from 2 to 6. One additional (RCT) study reported birth weight outcomes and study quality was assessed using the Jadad Scale. In some cases, studies had more than one comparison group, which consisted of groups of women who were drug-users and those who were not, of

which we used data from the comparison group of drug-users. Estimates of treatment effect were measured using the standardized mean difference. Results of the new weights incorporating study quality for birth weight gave a combined estimate with a 95% confidence interval of $d_{smd} = 0.2644$ (0.0860, 0.4428). We also combined studies using the traditional inverse variance weights using a traditional random effects model, giving a combined estimate and 95% confidence interval of $d_{smd} = 0.3032$ (-0.0725, 0.6788). Forest plots for each method of weighting are provided in Appendix A.

Results obtained from both weighting methods indicate similar point estimates for the combined estimate, but we find that incorporating study quality produced a statistically significant result indicating a small benefit for the treatment groups over the control groups. The result from the traditional weighting method reveals a slightly larger point estimate, but is associated with a confidence interval that has a lower limit below zero. A forest plot comparing the two methods of weighting is available in Appendix A. The combined estimate obtained from weighting by study quality can be interpreted as resulting from more accurate estimates of treatment effect from higher quality studies. We cannot conclude that the result from either weighting method is more accurate than the other. Taking into consideration the high degree of heterogeneity among studies reporting birth weight outcomes, we should not ignore the results obtained from incorporating study quality which suggests that integrated treatment programs have at least a small effect on birth weight outcomes.

4.4 Impact of Correlation Coefficients on Combined Estimates using Random Effects Meta-analysis – Dealing with Missing Data

We assessed the impact of correlation coefficients on overall combined estimates of the standardized mean gain d_{smg} for maternal depression measured by overall mood. There were three studies that reported pre-post measures of overall mood, in which all patients were followed-up at 6 months. The results of the overall combined estimates d_{smg} were 0.67 (-0.10, 1.45), 0.67 (0.04, 1.3), 0.67 (0.12, 1.2), 0.67 (0.18, 1.15) and 0.66 (0.25, 1.07) for corresponding values of $r = 0.2, 0.5, 0.65, 0.75$ and 0.85. A forest plot displaying the impact of the choice of r on combined estimates is available in Appendix A. Looking at the combined estimates of d_{smg} we notice the similarity between the point estimates, indicating the robustness of the results regardless of the choice of correlation coefficient. This is due to the fact that the correlation only affects the inverse variance weight, and not the effect size [32]. Thus, for all considered values of r the estimate of the combined effect size for overall mood is fairly large as a result of treatment. The confidence intervals on the other hand were affected by the choice of r, and must be interpreted with caution. As r increased, the width of the confidence intervals became narrower. Choosing a value of $r = 0.5$ for our final analysis allows us to incorporate a moderate correlation between pre and post assessments, also giving a robust overall combined estimate with a confidence interval that has a positive lower bound.

Chapter 5

Discussion

5.1 Summary of Key Findings

5.1.1 Combining *p*-values

In this study, I applied the *p*-value method for combining data for outcome metrics that were too dissimilar to be combined in the traditional way. Combining *p*-values is an approach adopted mainly due to its wide applicability [21], where *p*-values from diverse tests may be combined (see Section 3.2). In some cases, the *p*-value method may be the most appropriate method for combining data, though it has its limitations where, in general, it provides us with less detailed information. Since *p*-values incorporate a measure of precision, through sample size or degrees of freedom, small *p*-values may result partly because of large effects or sample sizes or due to experimental-design. For these reasons, traditional methods of combining estimates of treatment effect are generally preferred over combining *p*-values when appropriate [21].

Results of combining *p*-values for AAPI and Apgar scores do not allow us to draw conclusions about the size of the overall effect. Rather, it only tells us if at least some of the population experienced a significant effect as a result of integrated treatment. For four out of five AAPI subscales, we can conclude that at least some of the women experience a change in parenting attitudes from baseline. For Apgar scores, we can conclude that newborns in the treatment groups did not experience any significant effect when compared to the control groups. The use of the two-sided test is common in

primary research [28], and is the method we adopted due to the way data was collected and reported. Its use in combining *p*-values is problematic since it does not convey information about the direction of deviation from the null hypothesis. Thus, we are only able to conclude whether integrated treatment programs have any effect on outcomes or not. The evidence from combining *p*-values does not provide as much information as combining magnitudes of treatment effect, but still contributes valuable results that could be used for enhancing future research in substance use treatment for pregnant and parenting women.

5.1.2 Capture Re-capture Method

The results from the capture re-capture method show that we captured 90% of the horizon. Since an exhaustive literature search was conducted, a reasonably high capture rate would suggest that we have retrieved a sufficient amount of relevant studies. Although we did not plan to estimate the horizon in the protocol stage, it would be beneficial for future studies to consider using the capture re-capture technique in the design phase. In doing this, investigators can decide whether estimation of the horizon is appropriate and feasible for their study, and maintain a complete history of the capture re-capture patterns. In our meta-analysis we could not include hand-searching in our modeling step since we did not track the capture re-capture history. This method is also useful for investigators with limited resources and time, where capture re-capture methods can be used as a stopping rule for literature searching [19].

5.1.3 Assessing the Impact of Study Quality

Incorporating study quality as done by weighting can be thought of as a special case of exploring heterogeneity in a meta-analysis [20]. It allows us to explore if variation in study quality scores between studies could explain the variation in estimates of treatment effect. Since study quality is thought to be a moderator of treatment outcome, it is important to investigate its potential effect on combined estimates of treatment effect. We decided to incorporate study quality by weighting because there was not enough power to run meta-regression or subgroup analysis to explain heterogeneity.

Ignoring study quality may lead to unreliable results, since low quality studies are likely to produce erroneous estimates of treatment effect [33]. Poor quality studies may lead to bias in estimating treatment effects, or contribute extra variation when combined with high quality studies. For these reasons, meta-analyses combining studies with varying study quality may be susceptible to type I and II error due to biased estimates, as well as type II error from imprecise estimates [33]. It has been argued that because quality scores are not direct measures of precision, using them in the weighting of individual studies lacks statistical or empirical justification [20]. Another possible approach to incorporating study quality can be done by excluding studies of the poorest quality [34]. This can be thought of as an extreme case of weighting, where poor quality studies are given a weight of zero. This method was not appropriate for this meta-analysis because only a small number of studies were similar enough to be combined over each outcome and study quality scores were generally low.

It has been suggested that the best way for dealing with study quality in a meta-analysis is by conducting a sensitivity analysis [20]. In this way, we are able to assess the

robustness of the results from weighting by study quality. The results from Section 4.3 indicate similar point estimates for both weighting methods, but incorporating quality score into weights leads to a statistically significant result over the traditional weighting method. Although there is no consensus on how study quality should be handled in meta-analyses [20], it is obvious that it should be explored to encourage better quality studies in the future.

5.2 Comparison of Findings with Similar Works

Systematic reviews in substance abuse treatment to date have failed to include integrated treatment programs, and have primarily reported on evaluation of outcomes related to women's substance use, neglecting child and parenting outcomes [11-12]. A meta-analysis by Orwin *et. al.* (2001) [12] was conducted to assess the effectiveness of substance use treatment for women, in which three different group contrasts were examined: treatment vs. no treatment, women only vs. mixed gender programs and enhanced vs. standard women's treatment programs. Although they reported a handful of similar outcomes for women, including substance use, maternal well-being and pregnancy, they were not comparable as they did not specifically address integrated services for women and their children. The Orwin meta-analysis indicated similar limitations faced in this study, suggesting that substance abuse treatment studies are challenged by small number of subjects, variation in outcome measures, deficiencies in reporting and inconsistencies in timing of assessments.

5.3 Key Limitations of the Study

The purpose of this meta-analysis and systematic review was to address a broad spectrum of potential integrated treatment outcomes to address key issues missing in systematic reviews of substance use treatment studies to date. Though there has been an increase in research on integrated treatment programs, the majority of the studies conducted are observational. Single cohort designs are common as a result of internal program evaluations, where data are not collected primarily for research purposes. The lack of more rigorous designs such as RCTs may also result from ethical considerations, since there has been some evidence in the literature to indicate some benefits of integrated treatment, it would be unethical to refuse treatment to some patients.

Individual ratings of study quality in this meta-analysis were generally low, with a mean score of 1.4 (range 1-3) on the Jadad Scale, and 2.5 (range 0-6) on the NOS. A major reason for the poor study quality scores was due to the lack of reporting. It is important to note that poor reporting does not necessarily mean poor conduct of a study [35]. If items on the NOS or Jadad Scale were not reported, we could not assess the quality, and assumed that not reported translated to not done – although this assumption is not always accurate [36]. The Cochrane Collaboration [37] recommends the use of the CONSORT statement as a guideline for reporting on RCTs [37]. The CONSORT statement is a 22-item checklist that was developed to aid in the completeness and transparency of reporting of RCTs [38]. The STROBE statement is a similar checklist for observational studies that is based on the CONSORT statement [39]. Adhering to the CONSORT and STROBE statements can improve the overall quality of reporting in clinical studies. Other similar checklists for assessing the quality of reporting are also

available [40]. Improvement in the quality of reporting in clinical research will allow us to better assess the methodological quality of individual studies and determine the strength of evidence.

There was considerable heterogeneity in the studies in this meta-analysis in terms of outcome measures used, study design and timing of assessments. Taking into account all of these factors, there were few studies for any particular outcome that could be combined. For example, a primary outcome such as parenting could consist of attitude measures, involvement with child protection, custody status or attachment scores. Parenting attitudes alone could be measured using the AAPI subscales. Combining within each subscale and taking into account the variation in the timing of assessments and study design, only a few studies could be combined. When aggregating data across each of the outcomes, we came across the same issue in which we ended up with only a few studies that were similar enough to combine. In some cases, it resulted in single studies in which outcome measures could not be combined with those from any other study. For the studies that were combined, there was not enough power to conduct subsequent analyses such as exploring moderators of treatment outcome through meta-regression or subgroup analysis.

Missing data was also another limitation of this meta-analysis. Attempts were made at minimizing the amount of missing data through contacting authors for additional information. Data that were missing included effect sizes, information to calculate variances and other study variables. The correlation coefficient (r) needed to estimate the variance for the standardized mean gain played a large role in missing data since it was not reported in any of the cohort studies. To handle missing data of this type, a sensitivity

analysis was conducted to evaluate the impact of different values of r on combined estimate and its associated confidence interval. Although we did not impute missing effect sizes and other study level data in this thesis, the multiple imputation method [41] would be appropriate for assessing the impact of missing data on overall effects and inferences. A recent systematic review shows that missing data is a real problem in meta-analysis and requires a standardized approach to address it [42].

5.4 Implications for Clinical Practice and Further Research

Challenges faced in conducting this meta-analysis allow us to identify gaps in the research on integrated treatment programs for pregnant and parenting women and their children, such as poor reporting, small sample sizes, heterogeneity in measuring outcomes and timing of assessments. This indicates a need for researchers in the field to conduct larger studies of high quality and standardize reporting of related outcomes. It has been discussed that meta-analysis is used less in substance use treatment in comparison to related fields [12]. Higher quality evidence from individual trials will prove beneficial to future meta-analyses conducted in this field. Without large, high quality RCTs, it is hard to establish good evidence-based guidelines to care for substance-using women and their children. Research shows that meta-analysis of small underpowered trials can over- or under-estimate the benefits or risks/harms of interventions [43].

Poor quality of reporting is not necessarily worse in substance use literature [12], but the use of checklists and guidelines for minimum reporting such as the CONSORT and STROBE statements can be adopted to alleviate reporting issues. Both checklists are

recommended by the Cochrane Collaboration in the reporting of RCTs and observational studies [37], as well as the EQUATOR Network – a resource centre for promoting transparency and accuracy of reporting in health research [40]. Few journals in the area of substance use treatment endorse the CONSORT statement, while no related journals actively endorse the STROBE statement in their instructions for authors [44-45].

In a further attempt to promote systematic reviews and meta-analyses in this field, it is important to ensure that all relevant studies are retrieved and assessed for inclusion/exclusion criteria. This can be achieved through estimation of the horizon, which is not a requirement in standardized reporting guidelines for systematic reviews and meta-analyses, such as the QUOROM [46] and MOOSE statements [47]. The use of the capture re-capture technique for estimating the missing number of articles can be helpful in assessing the completeness of the current search, as well as identifying key sources for future literature searches. Software for estimating the horizon should be made more accessible to clinicians to promote ease and importance of its use in clinical research. In addition to searching relevant databases and journals, there is a substantial amount of unpublished literature in the area of integrated treatment programs for pregnant and parenting women. Future works on systematic reviews and meta-analyses should attempt to locate such fugitive literature through contacting relevant agencies, treatment programs and researchers.

Chapter 6

Conclusion

Overall, the results of combining data using the *p*-value method indicate some benefits of integrated treatment programs, while conducting random effects meta-analysis using traditional weights and weights incorporating study quality gave significantly different results. Through the use of the capture re-capture technique for estimation of the horizon, we are confident that we retrieved a high proportion of the relevant world literature for this meta-analysis.

For parenting attitudes measured by the AAPI subscales, results of the *p*-value method indicate evidence of benefit from treatment but stronger conclusions on the size of treatment effect could not be made. Combined *p*-values for neonatal behavioural outcomes measured by Apgar scores, on the other hand, did not provide us with sufficient evidence to believe that there was any effect from treatment. Results of the *p*-value method should be interpreted with caution, but despite its limitations, proper use of its results will enable better conduct of future research.

For birth weight outcomes, the results from incorporating study quality into weights and traditional inverse variance weights in a random effects meta-analysis did not agree. Since there was a high degree of heterogeneity in study quality, we may be inclined to believe the results from study quality weighting – that integrated treatment programs have a small, but positive effect on birth weight outcomes. In general,

comparing results from both weighting methods will give greater confidence in the results, or provide better insight into the interpretation if results are different.

For maternal well-being outcomes measured by overall mood (maternal depression), a sensitivity analysis of correlation coefficients indicated robust point estimates with wider confidence intervals as the value of r decreased. Since we expect outcomes between two time points to be correlated, our final choice of $r = 0.5$ gives a large effect on overall mood, with a confidence interval that has a positive lower limit.

Results from this thesis do not allow us to draw definitive conclusions of benefit or harm from integrated treatment programs because (i) we were limited by poor quality of evidence from individual studies and (ii) we did not consider all primary outcomes in this thesis. This thesis provides methods for dealing with common statistical issues in meta-analyses, particularly in areas of research where many of the studies are observational. The use of such statistical methods is important to incorporate into meta-analyses in order to thoroughly answer research questions and direct better quality research in the future.

References

1. Covington CY, Nordstrom-Klee B, Ager J, Sokol R, Delaney-Black V. (2002). Birth to age 7 growth of children prenatally exposed to drugs: A prospective cohort study. *Neurotoxicology & Teratology*, 24: 489-496.
2. Legrand LN, Iacono WG, McGue M. (2005). Predicting addiction: Behavioral genetics uses twins and time to decipher the origins of addiction and learn who is most vulnerable. *American Scientist*, 93: 140-147.
3. Kelley SJ. (1998). Stress and coping behaviors of substance-abusing mothers. *Journal of Social and Pediatric Nurses*, 3: 103-110.
4. Hans SL. (1999). Demographic and psychosocial characteristics of substance-abusing pregnant women. *Clinics in Perinatology*, 26: 55-74.
5. Horrigan TJ, Schroeder AV, Shaffer RM. (2000). The triad of substance abuse, violence, and depression are interrelated in pregnancy. *Journal of Substance Abuse Treatment*, 18: 55-58.
6. Curet LB, His HC. (2002). Drug abuse during pregnancy. *Clinical Obstetrics and Gynaecology*, 45: 73-88.
7. Hankin J, McCaul ME, Heussner J. (2000). Pregnant, alcohol-abusing women. *Alcoholism: Clinical and Experimental Research*, 24: 1276-1286.
8. Howell EM, Chasnoff IJ. (1999). Perinatal substance abuse treatment: findings from focus groups with clients and providers. *Journal of Substance Abuse Treatment*, 17: 139-148.

9. Niccols A, Sword W. (2005). "New Choices" for substance using mothers and their young children: Preliminary evaluation. *Journal of Substance Use*, 10 (4): 239-251.
10. Uziel-Miller ND, Lyons JS, Kissiel C, Love S. (1998). Treatment needs and initial outcomes of a residential recovery programs for African American women and their children. *American Journal of Addiction*, 7: 43-50.
11. Ashley OS, Marsden ME, Brady TM. (2003). Effectiveness of substance abuse treatment programming for women: A review. *American Journal of Drug and Alcohol Abuse*, 29: 19-53.
12. Orwin RG, Francisco L, Bernichon T. (2001). *Effectiveness of women's substance abuse treatment programs: A meta-analysis*. Arlington, VA: Battelle Centers for Public Health Research and Evaluation.
13. Belcher HME, Butz AM, Hoon AH, Reeves SA, Pulsifer MB. (2005). Spectrum of early intervention services for children with intrauterine drug exposure. *Infants & Young Children*, 18(1): 2-15.
14. Kern JK, West EY, Grannemann BD, Greer TL, Sness LM, Cline LL, *et al.* (2004). Reductions in stress and depressive symptoms in mothers of substance-exposed infants, participating in a psychosocial program. *Maternal and Child Health Journal*, 8(3):127-136.
15. Suchman N, Mayes L, Conti J, Slade A, Rounsaville B. (2004). Rethinking parenting interventions for drug-dependent mothers: From behaviour management to fostering emotional bonds. *Journal of Substance Abuse Treatment*, 27: 179-185.
16. Hasselblad V. (1995). Meta-analysis of environmental health data. *Science of the Total Environment*, 160-161: 545-58.

17. Goldsmith CH, Haynes RB, Garg AX, McKibbon KA, Wilczynski NL, Kastner M, *et al.* Horizon estimation – what is the horizon for a nephrology journal subset? Presentation (5th Canadian Cochrane Symposium, Ottawa, ON, February 12-13, 2007. Available at http://www.ccnc.cochrane.org/en/events.html. Accessed October 2008.

18. Bennett DA, Latham NK, Stretton C, Anderson CS. (2004). Capture- re-capture is a potentially useful method for assessing publication bias. *Journal of Clinical Epidemiology*, 57: 349-57.

19. Kastner M, Straus SE, McKibbon KA, Goldsmith CH. (2008). The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews. *Journal of Clinical Epidemiology*. Article in press: accepted 3 June 2008.

20. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. (2000). Methods for Meta-analysis in Medical Research. Chichester, England: Wiley.

21. Becker BJ. (1994). Combining significance levels. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage Foundation; 215-230.

22. George, EO. (1977). Combining independent one-sided and two-sided statistical tests: Some theory and applications. Unpublished doctoral dissertation, University of Rochester.

23. American Academy of Pediatrics, Committee on Fetus and Newborn, American College of Obstetricians and Gynecologists and Committee on Obstetric Practice. (2006). The Apgar Score. *Pediatrics*, 117: 1444-7.

24. Pecora PJ, Fraser MW, Nelson KE, McCroskey J, Meezan W. (1995). Evaluating Family-Based Services. New York: Aldine Transaction.

25. Wickens TD. (1989). Multiway Contingency Tables Analysis for the Social Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

26. Dobson, AJ. (2002). An Introduction to Generalized Linear Models, Second Edition. London: Chapman & Hall/CRC.

27. Foster, G, Goldsmith CH. Horizon estimation in SAS. (Paper 228-2008) SAS Global Forum Proceedings, San Antonio TX, 2008-03-16/19. http://www2.sas.com/proceedings/forum2008/228-2008.pdf.

28. Whitehead A. (2002). Meta-analysis of controlled clinical trials. West Sussex, England: John Wiley and Sons.

29. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed October 2008.

30. Moher, D, Jadad, AR, & Tugwell, P. (1996). Assessing the quality of randomized controlled trials. *International Journal of Technological Assessment of Health Care*, 12, 195-208.

31. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. (2008). Scales to Assess the Quality of Randomized Controlled Trials: A Systematic Review. *Physical Therapy*, 88: 156-75.

32. Lipsey MW, Wilson DB. (2001). Practical meta-analysis. Thousand Oaks, California: Sage Publications, Inc.

33. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology*, 45: 255-265.

34. Light RJ. (1987). Accumulating evidence from independent studies – what we can win and what we can lose. *Statistics in Medicine*, 6: 221-31.

35. Thabane L, Chu R, Cuddy K, Douketis J. (2007). What is the quality of reporting in weight loss intervention studies? A systematic review of randomized controlled trials. *International Journal of Obesity*, 31: 1554–1559.

36. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

37. The Cochrane Collaboration. Available at http://www.cochrane.org/. Accessed November 2008.

38. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, *et al.* (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, 276: 637-9.

39. von Elm E, Altman DG, Egger M, Pocock, SJ, Gøtzsche, Peter C, Vandenbroucke JP. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Epidemiology*, 18: 800-4.

40. Equator Network. Available at: http://www.equator-network.org/. Accessed November 2008.

41. Cheng J. (2007). A Systematic Review and Meta-analysis of Studies of Preoperative Aspirin on Bleeding and Cardiovascular Outcomes of Patients Undergoing Coronary Artery Bypass Surgery: A Comparison of Bayesian and Classical Approaches.

42. Wiebe N, Vandermeer B, Platt R, Klassen T, Moher D, Barrowman NJ. (2006). A systematic review identifies a lack of standardizations in methods for handling missing variance data. *Journal of Clinical Epidemiology*, 59: 342-53.

43. Furberg BD, Furberg CD. (2007). Evaluating Clinical Research: All that Glitters is not Gold, 2nd Edition. New York: Springer.

44. CONSORT. Available at http://www.consort-statement.org/. Accessed November 2008.

45. STROBE Statement. Available at http://www.strobe-statement.org/. Accessed November 2008.

46. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, 354:1896-900.

47. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. (2000). Meta-analysis of Observational Studies in Epidemiology. A Proposal for Reporting. *JAMA*, 283: 2008-12.

48. Conners NA, Bradley RH, Whiteside-Mansell L, Crone CC. (2001). A comprehensive substance abuse treatment program for women and their children: an initial evaluation. *Journal of Substance Abuse Treatment*, 21: 67-75.

49. McLellan AT, Kushner H, Metzger D, Peters R, Smith I, Grissom G, Pettinati H, Argeriou M. (1992). The fifth edition of the Addiction Severity Index. *Journal of Substance Abuse Treatment*, 9: 199-213.

50. Abidin RR. (1995). Parenting Stress Index Professional Manual, 3rd Edition. Odessa, FL: Psychological Assessment Resources, Inc.

51. Smith GR, Kramer T, Babor T, Burnam MA, Mosley CL, Rost K, Burns B. (1998). Substance Abuse Outcomes Module Users Manual. Available at http://www.netoutcomes.net.

52. Moos RH. (1974). Family environment scale. Palo Alto, CA: Consulting Psychologists Press.

53. Wilkinson GS. (1993). The Wide Range Achievement Test: Administration Manual. Wilmington, DE: Wide Range, Inc.

54. Bruininks RH, Woodcock RW, Weatherman RF, Hill BK. (1996). Scales of independent behavior-revised: comprehensive manual. Chicago, IL: Riverside Publishing Company.

55. Frankenburg WK, Dodds J. (1992). Denver II Training Manual, 2nd Edition. Denver, CO: Denver Developmental Material, Inc.

Appendix A

Figures

Figure 1.1: Graphical depiction of an integrated treatment program

Time Line	Intervention†
Baseline	a
Baseline to discharge	b c d e f g h i j k l m n
3, 6, and 12 months post discharge	Measurement of outcome†† Client outcomes: (1) Addiction Severity Index, 5th ed. (ASI) (2) Parenting Stress Index (PSI) (3) Outcomes of Addiction (4) Family Cohesiveness Scale (5) Intake/Follow-up demographic update form (6) The Wide Range Achievement Test III (WRAT III) Child outcomes: (7) Refusal skills (8) Adaptive behavior - Scales of Independent Behavior (SIB-R) (9) Denver Developmental Screening Test II (DDST II)

† Arkansas CARES, an integrated treatment program evaluated in a study by Conners *et. al.* [48].

†† Outcomes are described following this figure.

Figure 1.1: Graphical depiction of an integrated treatment program (continued)

a	A research assistant, who served as part of an independent evaluation team, conducted a baseline assessment one week after entry into the program. This consisted of in-person interviews with each client and paper-and-pencil questionnaires and developmental tests.
b	7-8 hour treatment day
c	Alcohol and drug abuse assessment, education, and treatment
d	Licensed mental health services for client, child, and family
e	Onsite residence
f	Licensed early intervention services
g	Onsite licensed childcare Infant and toddler Preschool and school-age School-age summer program
h	Parenting education and support
i	Health services and education
j	Community and onsite 12-step meetings
k	Life skills training
l	Group therapy

m	Case management
n	Group and individual counseling, covering the following areas: - Denial - The disease of addiction - Physiology and pharmacology of tobacco, alcohol, and other drugs - Effects of tobacco, alcohol, and other drugs on the mother and fetus - Parenting the drug-exposed infant - Women in relationships - Women's issues in recovery - Loss and grief - Family dynamics in recovery - 12-step recovery - Self-esteem building - Relapse prevention - AIDS and other sexually transmitted diseases - Family planning - Child development and care

Figure 1.2: Brief descriptions of outcomes from the study by Conners *et. al.* [48]:

(1) ASI	A semi-structured interview designed to gather information about aspects of a client's life which may contribute to their substance abuse problem. The ASI covers seven areas: medical, employment/support, alcohol, drug, legal, family/social, and psychiatric. [49]
(2) PSI	Gauges the stress the client perceives as it relates to parenting. The PSI consists of three scales (Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child). Each scale yields a score, and can be combined to achieve a total score. [50]
(3) The Outcomes of Addictions Questionnaire	Asks clients to indicate how frequently they have experienced various negative effects from drug or alcohol use in the four weeks prior to the assessment [51].
(4) Family Cohesiveness Scale	This form consists of nine items that are used to assess the degree to which the family acts as a single unit and expresses feelings of togetherness. The items were taken from the Family Environment Scale [52].
(5) Intake/Follow-up demographic update form	This instrument was developed by staff and evaluators to capture information about insurance, legal status, and the living situation of the client and her children [48].
(6) WRAT III	Designed to assess basic spelling, reading, and mathematics competence [53]. Scores on the WRAT III are converted to standard scores on a metric where the national mean is set at 100 and the SD is set at 15.
(7) Refusal skills	The child's ability to resist using drugs or alcohol is measured by a scale designed for this study. This instrument is completed by children eight years of age and older. [48]
(8) SIB-R	A comprehensive set of tests for measuring functional independence over a wide age range. Areas assessed include: motor development, daily living skills, social development, language, self-help skills, problem behaviors, and community adaptation. The SIB-R also provides an index of maladaptive behavior. [54]
(9) DDST II	An individually administered, norm-referenced, multi-skill device designed to assess developmental progress and to identify children with developmental problems [55].

Figure 2: Schema of study selection for analysis

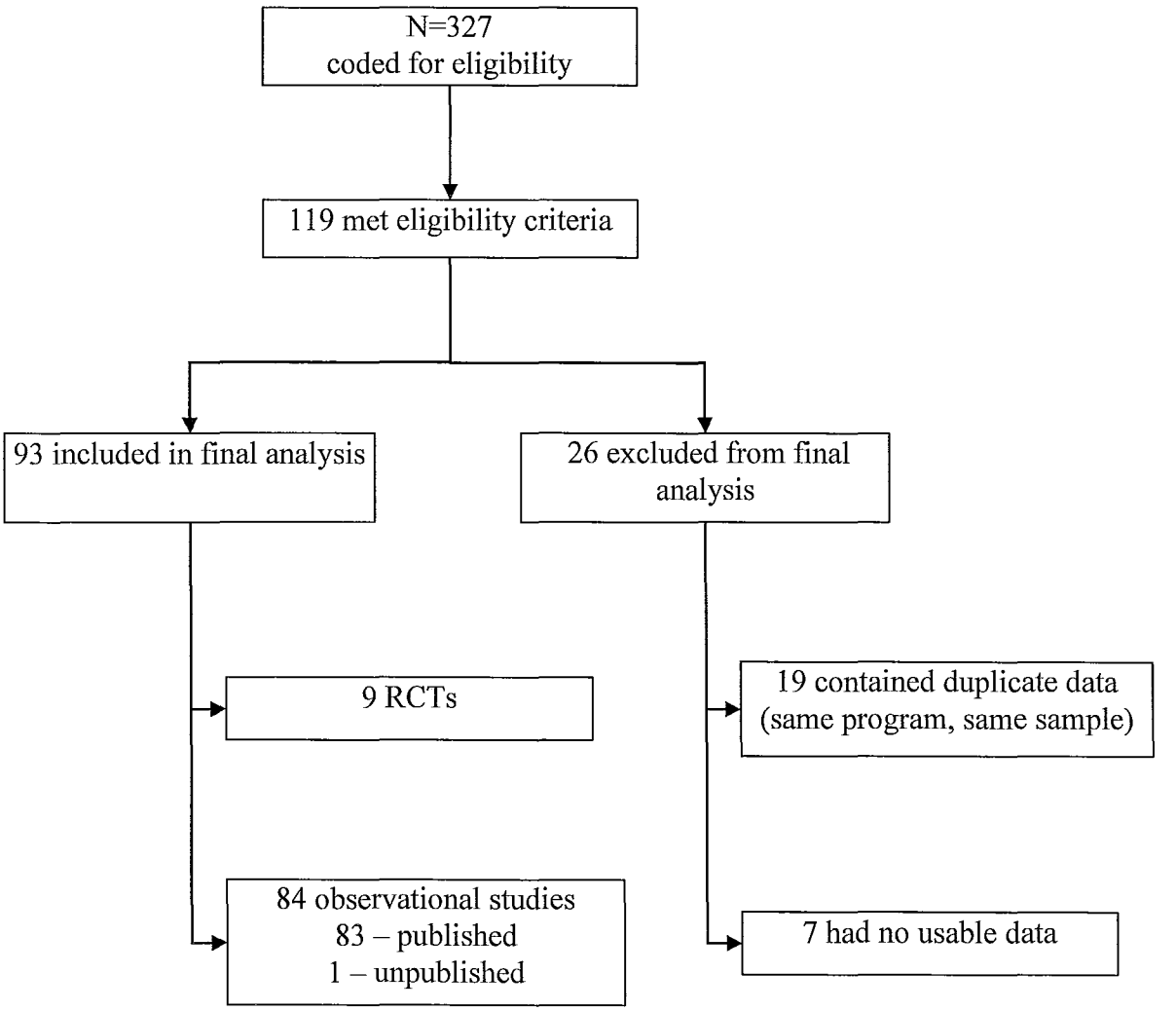
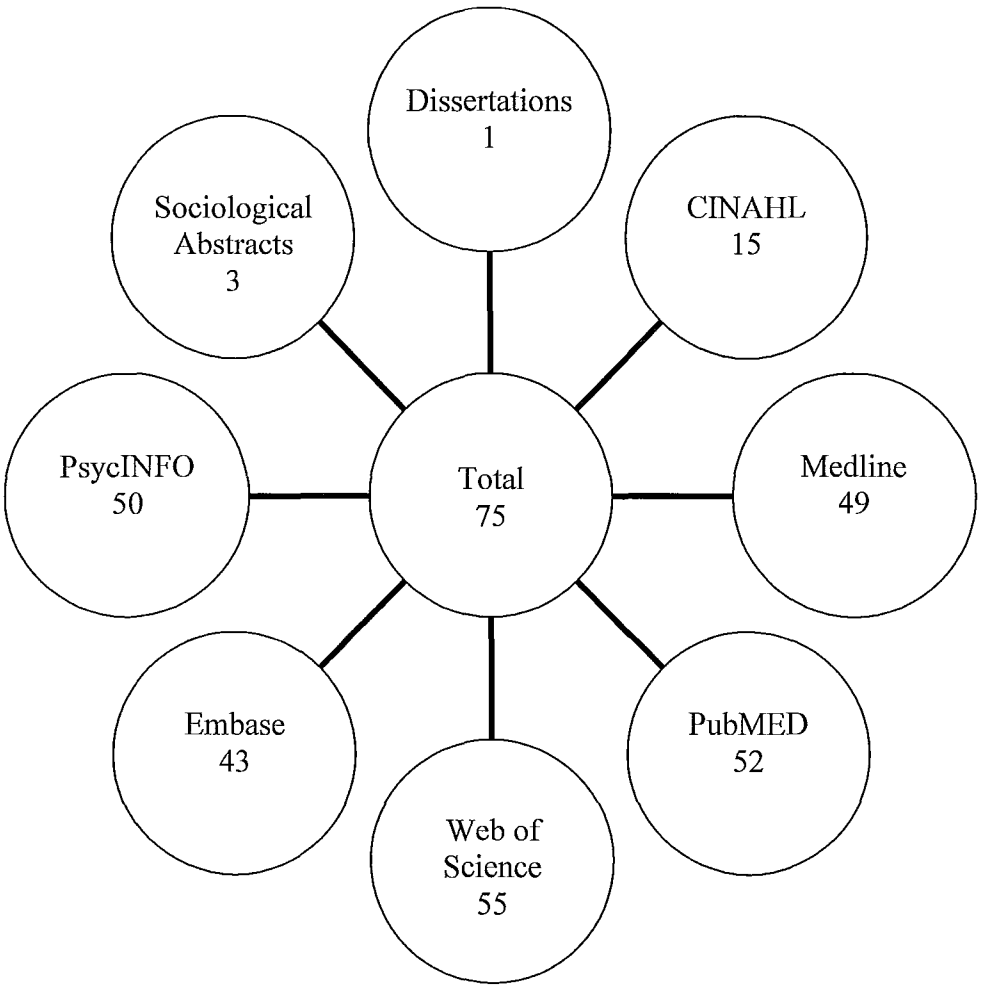


Figure 3: Graphical display of literature search



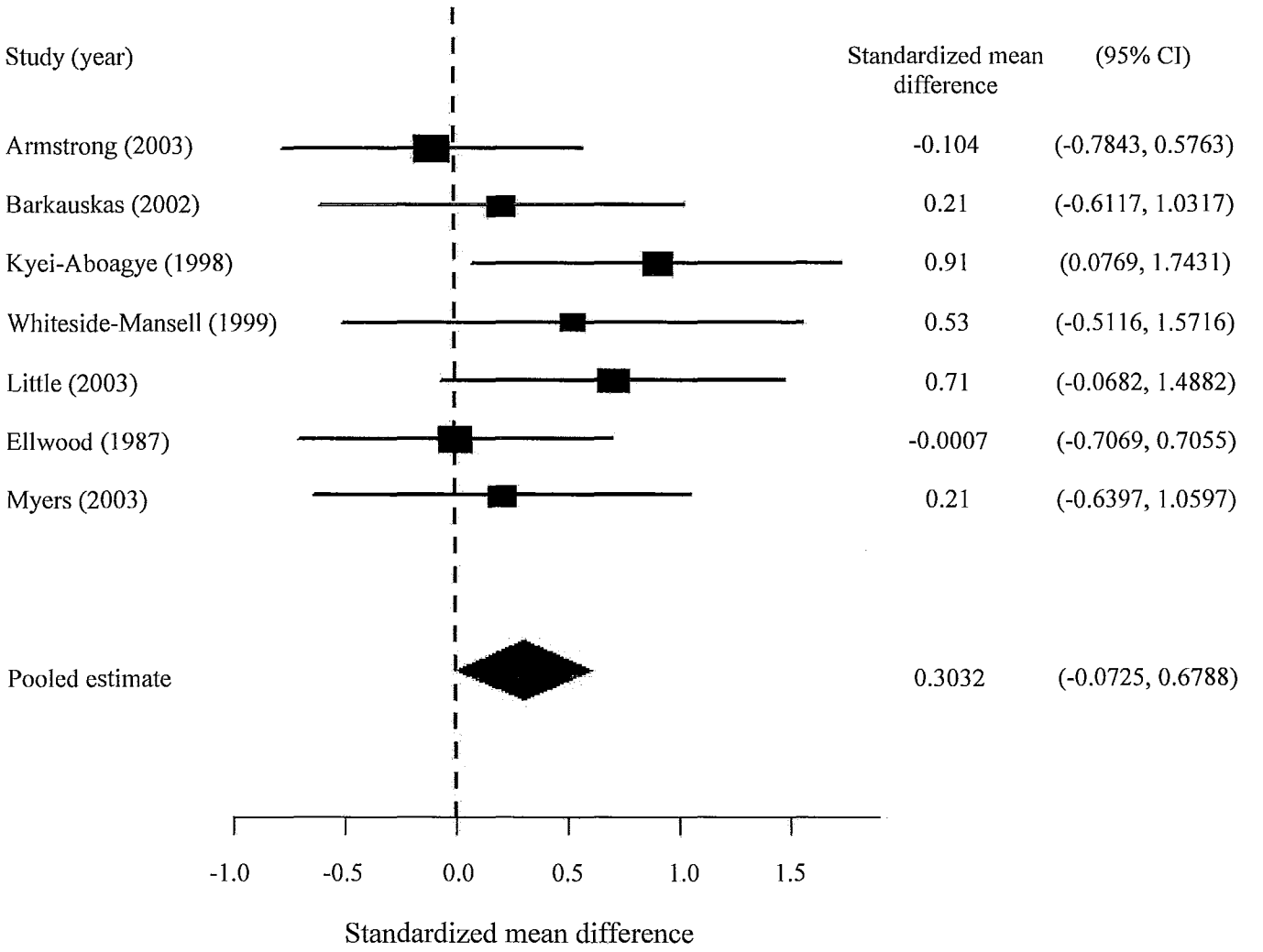


Figure 4: Classical random effects meta-analysis of birth weight outcomes

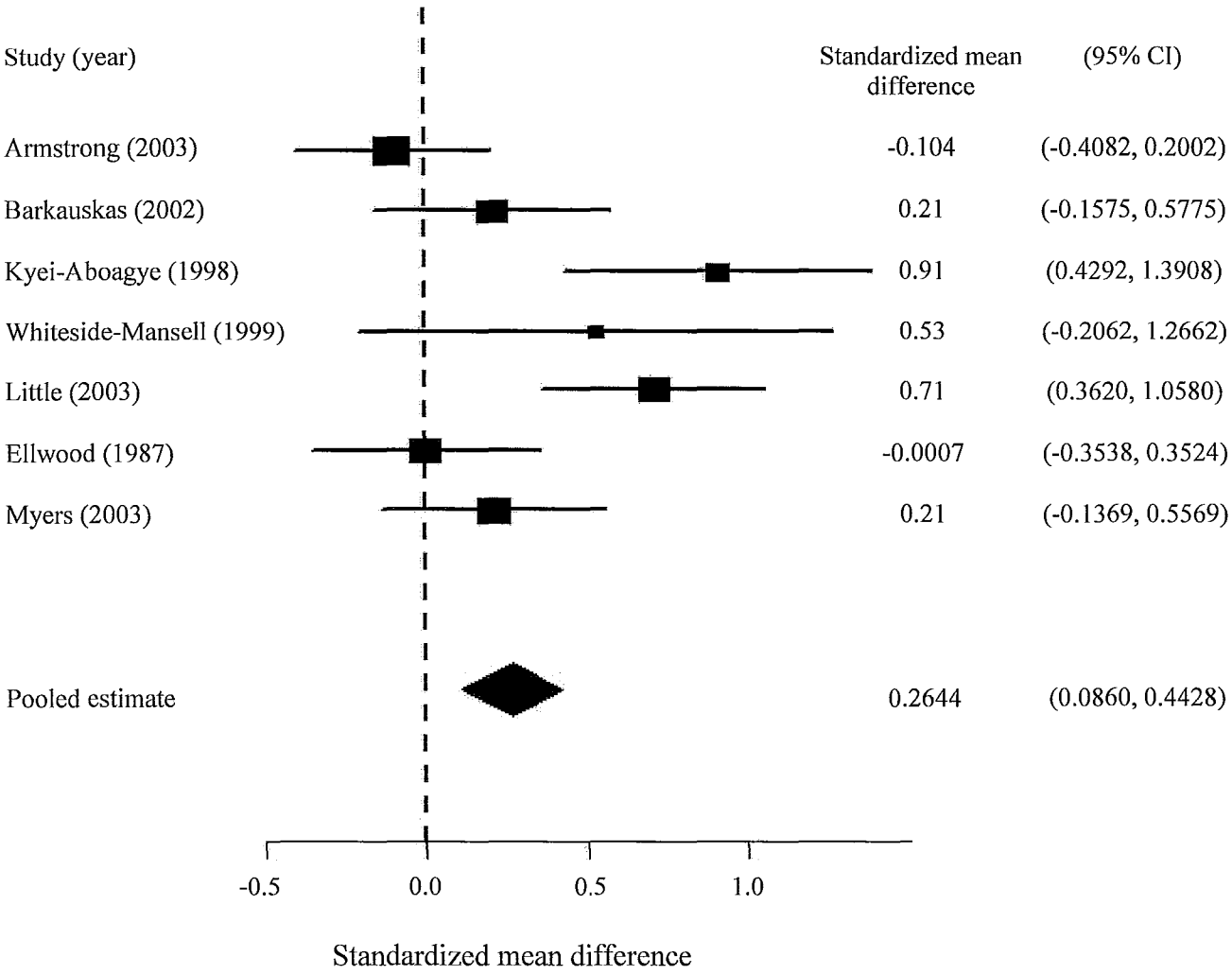


Figure 5: Meta-analysis using weights incorporating study quality of birth weight outcomes

Weighting Method	Overall combined estimate	(95% CI)
With SQ	0.2644	(0.0860, 0.4428)
Without SQ	0.3032	(-0.0725, 0.6787)

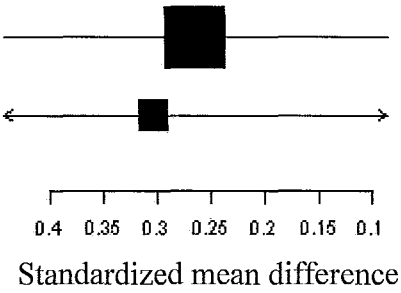


Figure 6: Comparison of incorporating study quality vs. inverse variance weighting in combining birth weight outcomes

Correlation Coefficient	Overall combined estimate	(95% CI)
0.2	0.67	(-0.1, 1.45)
0.5	0.67	(0.04, 1.3)
0.65	0.67	(0.12, 1.2)
0.75	0.67	(0.18, 1.15)
0.85	0.66	(0.25, 1.07)

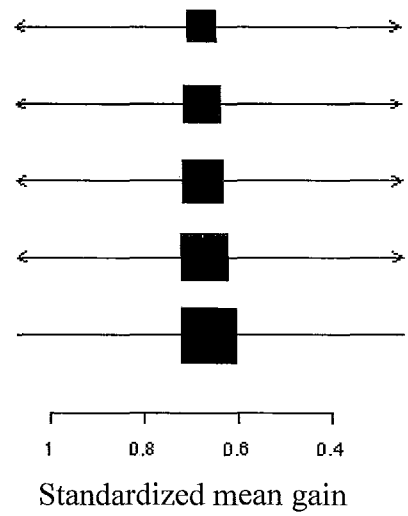


Figure 7: Impact of different correlation coefficients on overall estimates of maternal depression

Appendix B

Tables

Table 1: Description of selected outcomes for analysis

Outcome	Study design	Type of effect size
AAPI	Cohort	Standardized mean gain
Apgar scores	Cohort	Standardized mean gain
Birth weight	Quasi-experimental	Standardized mean difference
Overall mood	Cohort	Standardized mean gain

Table 2: Incomplete contingency table for literature search

row	PsycINFO	PubMED	Web of Science	Embase	count
1	1	1	1	1	26
2	1	1	1	0	5
3	1	1	0	1	1
4	1	1	0	0	1
5	1	0	1	1	4
6	1	0	1	0	5
7	1	0	0	1	4
8	1	0	0	0	5
9	0	1	1	1	7
10	0	1	1	0	4
11	0	1	0	1	0
12	0	1	0	0	8
13	0	0	1	1	1
14	0	0	1	0	4
15	0	0	0	1	0
16	0	0	0	0	.

Table 3: Results of the modeling step for estimating the horizon of the literature search

Estimating the Horizon
Comparison of Various Models

Obs	Model evaluated	Missing cell estimate	Missing lower cell 95% CI	Missing upper cell 95% CI	# of known items	Coefficient estimate	Coefficient lower 95% CI	Coefficient upper 95% CI	Horizon estimate	Horizon lower 95% CI	Horizon upper 95% CI	Known items as % of horizon	Deviance	DF
1	a b	11	4	22	70	2.3383	1.6089	3.0677	81	74	92	86%	0	0
2	a b c	3	1	6	71	1.0349	0.3289	1.7408	74	72	77	96%	17.68	3
3	a b c ab	1	0	6	71	-0.4055	-2.4229	1.612	72	71	77	99%	13.19	2
4	a b c ac	5	2	12	71	1.5847	0.7663	2.4032	76	73	83	93%	13.42	2
5	a b c bc	6	2	13	71	1.7405	0.9541	2.5268	77	73	84	92%	9.32	2
6	a b c ab ac	2	0	11	71	0.2231	-1.9461	2.3924	73	71	82	97%	10.78	1
7	a b c ab bc	2	0	16	71	0.539	-1.6313	2.7093	73	71	87	97%	7.43	1
8	a b c ac bc	20	6	64	71	2.9957	1.8362	4.1553	91	77	135	78%	0.46	1
9	a b c ab ac bc	10	0	121	71	2.2662	-0.2551	4.7876	81	71	192	88%	0	0
10	a b c d	1	0	2	75	-0.1165	-0.8456	0.6125	76	75	77	99%	47.01	10
11	a b c d ab	1	0	2	75	-0.4612	-1.4848	0.5625	76	75	77	99%	45.86	9
12	a b c d ac	2	0	3	75	0.3076	-0.462	1.0772	77	75	78	97%	41.74	9
13	a b c d ad	2	0	4	75	0.4327	-0.3867	1.2522	77	75	79	97%	42.76	9
14	a b c d bc	2	0	4	75	0.407	-0.3451	1.1591	77	75	79	97%	37.62	9
15	a b c d bd	2	0	4	75	0.2553	-0.5992	1.1098	77	75	79	97%	45.22	9
16	a b c d cd	2	0	4	75	0.5263	-0.2226	1.2752	77	75	79	97%	35.53	9
17	a b c d ab ac	1	0	3	75	-0.0171	-1.0828	1.0486	76	75	78	99%	40.81	8
18	a b c d ab ad	2	0	4	75	0.11	-1.0057	1.2256	77	75	79	97%	41.91	8
19	a b c d ab bc	2	0	4	75	0.0925	-0.9607	1.1456	77	75	79	97%	36.76	8
20	a b c d ab bd	1	0	3	75	-0.085	-1.2252	1.0551	76	75	78	99%	44.25	8

a = PubMED, b = PsycINFO, c = Embase, d = Web of Science

Estimating the Horizon

Comparison of Various Models

Obs	Model evaluated	Missing cell estimate	Missing lower cell 95% CI	Missing upper cell 95% CI	# of known items	Coefficient estimate	Coefficient lower 95% CI	Coefficient upper 95% CI	Horizon estimate	Horizon lower 95% CI	Horizon upper 95% CI	Known items as % of horizon	Deviance	DF
21	a b c d ab cd	2	0	4	75	0.2231	-0.831	1.2773	77	75	79	97%	34.75	8
22	a b c d ac ad	3	1	7	75	0.9428	0.0648	1.8208	78	76	82	96%	36.83	8
23	a b c d ac bc	3	1	6	75	0.8903	0.0842	1.6963	78	76	81	96%	31.74	8
24	a b c d ac bd	3	0	6	75	0.7482	-0.1612	1.6576	78	75	81	96%	39.54	8
25	a b c d ac cd	3	1	7	75	1.0264	0.2206	1.8321	78	76	82	96%	29.46	8
26	a b c d ad bc	3	1	7	75	1.0578	0.1954	1.9202	78	76	82	96%	32.51	8
27	a b c d ad bd	3	0	7	75	0.9164	-0.0529	1.8856	78	75	82	96%	40.38	8
28	a b c d ad cd	4	1	8	75	1.2073	0.3437	2.0709	79	76	83	95%	30.14	8
29	a b c d bc bd	3	0	6	75	0.8621	-0.0323	1.7565	78	75	81	96%	35.29	8
30	a b c d bc cd	4	1	7	75	1.1361	0.3467	1.9254	79	76	82	95%	25.11	8
31	a b c d bd cd	3	1	7	75	1.0069	0.1113	1.9025	78	76	82	96%	33.02	8
32	a b c d ab ac ad	2	0	7	75	0.6678	-0.5379	1.8735	77	75	82	97%	36.36	7
33	a b c d ab ac bc	2	0	5	75	0.4513	-0.5812	1.4839	77	75	80	97%	28.45	7
34	a b c d ab ac bd	2	0	6	75	0.439	-0.7801	1.6581	77	75	81	97%	38.91	7
35	a b c d ab ac cd	3	0	7	75	0.7825	-0.3539	1.9189	78	75	82	96%	29.07	7
36	a b c d ab ad bc	3	0	8	75	0.8074	-0.395	2.0099	78	75	83	96%	32.14	7
37	a b c d ab ad bd	2	0	6	75	0.6237	-0.537	1.7844	77	75	81	97%	38.94	7
38	a b c d ab ad cd	3	0	10	75	0.9915	-0.2338	2.2169	78	75	85	96%	29.89	7
39	a b c d ab bc bd	2	0	6	75	0.5709	-0.6442	1.7859	77	75	81	97%	34.76	7
40	a b c d ab bc cd	3	0	8	75	0.9193	-0.2096	2.0482	78	75	83	96%	24.81	7
41	a b c d ab bd cd	3	0	8	75	0.7388	-0.4943	1.9719	78	75	83	96%	32.61	7
42	a b c d ac ad bc	6	2	17	75	1.7871	0.7918	2.7824	81	77	92	93%	24.98	7

a = PubMED, b = PsycINFO, c = Embase, d = Web of Science

Estimating the Horizon

Comparison of Various Models

Obs	Model evaluated	Missing cell estimate	Missing lower cell 95% CI	Missing upper cell 95% CI	# of known items	Coefficient estimate	Coefficient lower 95% CI	Coefficient upper 95% CI	Horizon estimate	Horizon lower 95% CI	Horizon upper 95% CI	Known items as % of horizon	Deviance	DF
43	a b c d ac ad bd	6	1	17	75	1.6875	0.5506	2.8244	81	76	92	93%	33.22	7
44	a b c d ac ad cd	4	1	10	75	1.3605	0.5115	2.2096	79	76	85	95%	26.91	7
45	a b c d ac bc bd	5	1	13	75	1.5319	0.5221	2.5417	80	76	88	94%	28.38	7
46	a b c d ac bc cd	7	2	16	75	1.8274	0.9233	2.7316	82	77	91	91%	17.38	7
47	a b c d ac bd cd	6	2	17	75	1.7449	0.7111	2.7787	81	77	92	93%	25.64	7
48	a b c d ad bc bd	7	2	21	75	1.8728	0.7295	3.016	82	77	96	91%	28.46	7
49	a b c d ad bc cd	10	3	26	75	2.2033	1.1782	3.2283	85	78	101	88%	16.94	7
50	a b c d ad bd cd	9	2	31	75	2.1858	0.9684	3.4032	84	77	106	89%	25.3	7
51	a b c d bc bd cd	4	1	9	75	1.247	0.3787	2.1154	79	76	84	95%	24.83	7
52	a b c d ab ac ad bc	4	1	13	75	1.322	0.0854	2.5586	79	76	88	95%	23.04	6
53	a b c d ab ac ad bd	5	1	16	75	1.41	0.0706	2.7494	80	76	91	94%	32.45	6
54	a b c d ab ac ad cd	4	0	12	75	1.1954	-0.0409	2.4318	79	75	87	95%	26.77	6
55	a b c d ab ac bc bd	3	0	10	75	1.0405	-0.1988	2.2798	78	75	85	96%	25.98	6
56	a b c d ab ac bc cd	5	1	13	75	1.4163	0.276	2.5565	80	76	88	94%	15.56	6
57	a b c d ab ac bd cd	5	1	20	75	1.5908	0.1915	2.9902	80	76	95	94%	25.53	6
58	a b c d ab ad bc bd	6	1	20	75	1.6198	0.2674	2.9723	81	76	95	93%	27.88	6
59	a b c d ab ad bc cd	9	2	37	75	2.1688	0.7505	3.5872	84	77	112	89%	16.93	6
60	a b c d ab ad bd cd	8	1	31	75	1.9722	0.5229	3.4215	83	76	106	90%	24.98	6
61	a b c d ab bc bd cd	3	0	10	75	1.041	-0.1985	2.2805	78	75	85	96%	24.61	6
62	a b c d ac ad bc bd	21	5	82	75	3.0152	1.6293	4.4011	96	80	157	78%	18.82	6
63	a b c d ac ad bc cd	12	4	33	75	2.4583	1.4268	3.4899	87	79	108	86%	12.67	6
64	a b c d ac ad bd cd	13	3	45	75	2.5307	1.2618	3.7996	88	78	120	85%	21	6

a = PubMED, b = PsycINFO, c = Embase, d = Web of Science

Estimating the Horizon

Comparison of Various Models

Obs	Model evaluated	Missing cell estimate	Missing lower cell 95% CI	Missing upper cell 95% CI	# of known items	Coefficient estimate	Coefficient lower 95% CI	Coefficient upper 95% CI	Horizon estimate	Horizon lower 95% CI	Horizon upper 95% CI	Known items as % of horizon	Deviance	DF
65	a b c d ac bc bd cd	9	2	23	75	2.0949	1.0623	3.1276	84	77	98	89%	16.44	6
66	a b c d ad bc bd cd	16	4	58	75	2.7476	1.4493	4.046	91	79	133	82%	15.05	6
67	a b c d ab ac ad bc bd	15	3	71	75	2.6915	1.1303	4.2526	90	78	146	83%	16.42	5
68	a b c d ab ac ad bc cd	9	2	32	75	2.1057	0.7619	3.4495	84	77	107	89%	11.98	5
69	a b c d ab ac ad bd cd	12	2	52	75	2.4052	0.8744	3.936	87	77	127	86%	20.91	5
70	a b c d ab ac bc bd cd	6	1	20	75	1.6475	0.3414	2.9535	81	76	95	93%	15.07	5
71	a b c d ab ad bc bd cd	15	3	72	75	2.6986	1.1285	4.2688	90	78	147	83%	15.04	5
72	a b c d ac ad bc bd cd	26	6	99	75	3.2214	1.8568	4.5859	101	81	174	74%	9.73	5
73	a b c d ab ac ad bc bd cd	20	3	95	75	2.9478	1.3517	4.5439	95	78	170	79%	9.12	4
74	a b c d ac ad bc cd abc	12	4	33	75	2.4583	1.4268	3.4899	87	79	108	86%	12.2	5
75	a b c d ac ad bc cd abd	14	4	41	75	2.6221	1.5418	3.7024	89	79	116	84%	11.81	5
76	a b c d ac ad bc cd acd	8	2	24	75	1.9841	0.819	3.1493	83	77	99	90%	7.7	5
77	a b c d ac ad bc cd bcd	12	4	33	75	2.4583	1.4268	3.4899	87	79	108	86%	10.46	5
78	a b c d ac ad bc cd acd bcd	8	2	24	75	1.9841	0.819	3.1493	83	77	99	90%	5.49	4
79	a b c d ab ac ad bc bd cd abc acd	11	1	107	75	2.3582	0.0494	4.6671	86	76	182	87%	4.71	2
80	a b c d ab ac ad bc bd cd abc bcd	107	10	1127	75	4.6667	2.3065	7.0269	182	85	1202	41%	0.84	2
81	a b c d ab ac ad bc bd cd abd acd	1	0	1	75	-21.1052	-23.3915	-18.8189	76	75	76	99%	2.01	2
82	a b c d ab ac ad bc bd cd abd bcd	42	4	411	75	3.7233	1.4303	6.0163	117	79	486	64%	0.92	2
83	a b c d ab ac ad bc bd cd acd bcd	39	2	569	75	3.6601	0.9775	6.3427	114	77	644	66%	0	2

a = PubMED, b = PsycINFO, c = Embase, d = Web of Science

Appendix C

SAS code for analysis

SAS code for combining *p*-values for AAPI

```
/*CONVERTING PARENTING DATABASE FROM SPSS FORMAT TO SAS DATA SET*/

filename allaapi "C:\Documents and Settings\Jennifer Liu\My
Documents\MSc Thesis\Analysis\spss por files\all aapi_new_adj.por";
proc convert spss = allaapi out = aapi;
run;

/*FOR OUTCOME MEASURE: 1 AND STUDY DESIGN: COHORT */
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 1 then output;
run;

/*FOR OUTCOME: 2  (parenting attitudes - inappropriate expectations)
AND STUDY DESIGN: COHORT */
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 2 then output;
run;

/*FOR OUTCOME: 3 (parenting attitudes - lack of empathy) AND STUDY
DESIGN: COHORT */
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 3 then output;
run;

/*FOR OUTCOME: 4 (parenting attitudes - corporal punishment) AND STUDY
DESIGN: COHORT */
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 4 then output;
run;

/*FOR OUTCOME: 5 (parenting attitudes - role reversal) AND STUDY
DESIGN: COHORT */
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 5 then output;
run;

/*FOR OUTCOME: 50 (AAPI - power independence) AND STUDY DESIGN: COHORT
*/
data aapi1;
set aapi;
if DESIGN4A = 3 & ES_4AMSR = 50 then output;
run;

/*TAKE INTO ACCOUNT TIMING OF ASSESSMENT*/
data aapi2;
set aapi1;
```

```
/*FOR 3 MONTH FOLLOW-UP*/
if TIME_ASS = 1 | TIME_ASS = 7 | TIME_ASS = 21 | TIME_ASS = 24 then
output;
run;

/*FOR 6 MONTH */
if TIME_ASS = 1 | TIME_ASS = 7 | TIME_ASS = 14 | TIME_ASS = 16 |
TIME_ASS = 21 then output;
run;

/*FOR 12 MONTH */
if TIME_ASS = 1 | TIME_ASS = 7 | TIME_ASS = 15 | TIME_ASS = 17 |
TIME_ASS = 43  | TIME_ASS = 21 then output;
run;


/*CONVERT EFFECT SIZE TO P-VALUE*/

data aapi3;
set aapi2;
if P_VAL = . then P_VAL=2*probt(-ABS(ES_4A),NANALY4A+N_1-2);
run;

proc print data = aapi3 noobs;
var P_VAL;
run;

/*COMBINING P-VALUES USING LOGIT METHOD*/

data logit_aapi;
set aapi3;
z = log(P_VAL/(1-P_VAL)) ;
run;

proc summary data = logit_aapi;
var z AAPI;
output out = p_aapi sum = x k;
run;

data p_aapi1;
set p_aapi;
pi=constant('pi');
q=-x * ( 14*(pi**2)* ( (5*k+2)/(3*(5*k+4)) ) )**(-.5);
p=1-probt(q,5*k+4);
run;

proc print data = p_aapi1;
var p;
var q;
run;
```

SAS code for combining *p*-values for Apgar scores

```
/* CONVERTING AGPAR AND NBAS DATABASE FROM SPSS FORMAT TO SAS DATA
SET*/

filename ag "C:\Documents and Settings\Jennifer Liu\My Documents\MSc
Thesis\Analysis\spss por files\agpar & NBAS_new.por";
proc convert spss = ag out = agpar;
run;

/*FOR OUTCOME MEASURE: 1 MINUTE AGPAR AND STUDY DESIGN: QUASI-
EXPERIMENTAL*/

data agpar1;
set agpar;
if DESIGN5A = 2 & ES_5AMSR = 14 & GROUP = 2 then output;
if DESIGN5A = 2 & ES_5AMSR = 14 & GROUP = 4 then output;
run;

/*CONVERT EFFECT SIZES TO P-VALUES*/

data agpar2;
set agpar1;
if DATATP5A = 1 then P_VAL=2*probt(-ABS(ES_5A),NANALY5A+N_TX-2);
if DATATP5A = 11 then P_VAL = 2*probnorm(-ABS(Z_VAL));
run;

proc print data = agpar2 noobs;
var P_VAL;
run;

/*COMBINING P-VALUES USING LOGIT METHOD*/

data logit_agpar;
set agpar2;
z = log(P_VAL/(1-P_VAL)) ;
run;

proc summary data = logit_agpar;
var z agpar;
output out = p_agpar sum = x k;
run;

data p_agpar1;
set p_agpar;
pi=constant('pi');
q=-x * ( 14*(pi**2)* ( (5*k+2)/(3*(5*k+4)) ) )**(-.5);
p=1-probt(q,5*k+4);
run;

proc print data = p_agpar1;
var p;
var q;
run;
```

SAS code for weighting studies by traditional method and by incorporating study quality

```
/*random-effect meta-analysis using inverse variance as weight*/

data remlma;
set bw2;
var=1/INV_VAR;
STUDY_ID=_n_;
col=_n_;
row=_n_;
value = var;
run;

proc mixed data=remlma method =reml order=data;
class STUDY_ID;
model ES_5A =/ covb solution;
random STUDY_ID/gdata=remlma solution;
repeated diag;
run;


/*weighting by study quality X inverse variance as weight*/

data newbw;
set bw2;
tau2 = 0.119;
wstar = 1/(1/INV_VAR + tau2);
est = 1/(NEWCAST * wstar);
run;

proc mixed cl method=ml data = newbw;
class STUDY_ID;
model ES_5A = / s cl;
repeated /group=STUDY_ID;
parms/ parmsdata=newbw eqcons = 1 to 7;
run;
```

SAS code for combining *p*-values for Apgar scores

```
/* CONVERTING MATERNAL DEPRESSION DATABASE FROM SPSS FORMAT TO SAS DATA
SET*/

filename dep "C:\Documents and Settings\Jennifer Liu\My Documents\MSc
Thesis\Analysis\spss por files\Depression_new_adj.por";
proc convert spss = dep out = matdep;
run;

/*FOR MATERNAL DEPRESSION: OVERALL MOOD AND COHORT DESIGN */

data depscores1;
set matdep;
if DESIGN3A = 3 & ES_3AMSR = 10 then output;
run;

/*FOR 6 MONTH FOLLOW-UP */

data depscores;
set depscores1;
if TIME_ASS = 14 | TIME_ASS = 16 then output;
run;

data depscores2;
set depscores;
if COHEN = . then delete;
if DATATP3A NE 1 then delete;
run;

/*RANDOM-EFFECT FOR r1*/
proc mixed cl method=reml data = depscores3;
class STUDY_ID;
model COHEN = / s cl;
random int/ subject=STUDY_ID s;
repeated /group=STUDY_ID;
parms (0.01 to 2.00 by 0.01)
(.042802)(0.090178)(0.084802)
/eqcons=2 to 4;
run;

/*RANDOM-EFFECT FOR r2*/
proc mixed cl method=reml data = depscores3;
class STUDY_ID;
model COHEN = / s cl;
random int/ subject=STUDY_ID s;
repeated /group=STUDY_ID;
parms (0.01 to 2.00 by 0.01)
(0.031691)(0.070178)(0.064802)
/eqcons=2 to 4;
run;

/*RANDOM-EFFECT FOR r3*/
proc mixed cl method=reml data = depscores3;
class STUDY_ID;
model COHEN = / s ;
```

```
random int/ subject=STUDY_ID s;
repeated /group=STUDY_ID;
parms (0.01 to 2.00 by 0.01)
(0.024283)(0.056844)(0.051468)
/eqcons=2 to 4;
run;


/*RANDOM-EFFECT FOR r4*/
proc mixed cl method=reml data = depscores3;
class STUDY_ID;
model COHEN = / s cl;
random int/ subject=STUDY_ID s;
repeated /group=STUDY_ID;
parms (0.01 to 2.00 by 0.01)
(0.016876)(0.043511)(0.038135)
/eqcons=2 to 4;
run;

/*RANDOM-EFFECT FOR r5*/
proc mixed cl method=reml data = depscores3;
class STUDY_ID;
model COHEN = / s cl;
random int/ subject=STUDY_ID s;
repeated /group=STUDY_ID;
parms (0.01 to 2.00 by 0.01)
( 0.06502)( 0.13018)( 0.12480)
/eqcons=2 to 4;
run;
```