Therapist Variation within Meta-Analyses of Psychotherapy Trials

A thesis submitted to The University of Manchester for the degree of

DOCTOR OF PHILOSOPHY

in the Faculty of Medical and Human Sciences

2010

By Rebecca E. A. Walwyn

Biostatistics, Health Sciences Research Group,
School of Medicine

List of Contents

Word Count 71,392

List of Tables

List of Figures

List of Boxes

Abbreviations

ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AMD	Absolute Mean Difference
APT	Adaptive Pacing Therapy
BAC	British Association of Counselling
BACP	British Association of Counselling and Psychotherapy
BDI	Beck Depression Inventory
BSI	Brief Symptom Index
BT	Behaviour Therapy
CBT	Cognitive Behaviour Therapy
CCDAN	Cochrane Collaboration Depression, Anxiety and Neurosis
CDC	Centers for Diseases Control
CFS/ME	Chronic Fatigue Syndrome / Myalgic Encephalomyelitis or Encephalopathy
CI	Confidence Interval
CIS-R	Clinical Interview Schedule – Revised
CONSORT	Consolidated Standards of Reporting Trials
COREC	Central Office for Research Ethics Committees
CPN	Community Psychiatric Nurse
CTU	Clinical Trials Unit
D-L	DerSimonian-Laird
EPQ	Eysenck Personality Questionnaire
EWS	Early Warning Signs
FWA	Family Welfare Association
GET	Graded Exercise Therapy
GHQ	General Health Questionnaire
GP	General Practitioner
HADS-D	Hospital Anxiety and Depression Scale – Depression subscale
ICC	Intraclass Correlation Coefficient
IID	Independent and Identically Distributed
IIP-32	Inventory of Interpersonal Problems – 32
IMI-CM	Imipramine Hydrochloride plus Clinical Management
IPD	Individual Patient Data
IPT	Interpersonal Psychotherapy
LOCF	Last Observation Carried Forward
MANOVA	Multivariate Analysis of Variance
MCMC	Markov Chain Monte Carlo
MD	Mean Difference
MeSH	Medical Subject Headings
MH&N	Mental Health and Neuroscience
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator/Estimate
MRC	Medical Research Council
MS	Multiple Sclerosis
MST	Multisystemic Therapy
MSW	Medical Social Worker
NIMH	National Institute of Mental Health
NP	Nurse Practitioner
NRT	Nicotine Replacement Therapy

NS	Not Specified
OCD	Obsessive Compulsive Disorder
OR	Odds Ratio
PACE	Pacing, graded Activity and Cognitive behaviour therapy: a randomised Evaluation
PIAG	Patient Information Advisory Group
PLA-CM	Placebo plus Clinical Management
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PST	Psychosocial Treatment
PTSD	Post-Traumatic Stress Disorder
RCT	Randomised Controlled Trial
REML	Restricted Maximum Likelihood
RevMan	Review Manager
RIGLS	Restricted Iterative Generalised Least Squares
R&D	Research and Development
SAS-M	Social Adjustment Scale – Modified
SD	Standard Deviation
SE	Standard Error
SF-36	Short-Form – 36
SI	Symptom Index
SMD	Standardised Mean Difference
SSMC	Standardised Specialist Medical Care
STPP	Short-Term Psychodynamic Psychotherapy
SUPPORT	Supportive Psychotherapy
TAU	Treatment as Usual
TDCRP	Treatment for Depression Collaborative Research Program
UMVUE	Uniformly Minimum Variance Unbiased Estimator/Estimate

Notation

General

h	Study
i	Arm
j	Therapist
l	Patient

H	Total number of studies
f	Number of arms
k	Number of therapists
K	Total number of therapists
m	Number of patients per therapist, or cluster size
m_0	Average cluster size
$\overline{m}$	Arithmetic mean cluster size
$\hat{m}_1$	Harmonic mean cluster size
n	Number of patients, or sample size
N	Total number of patients, or sample size
R	Allocation ratio of patients across arms
ρ	Intraclass correlation coefficient
$deff$	Design effect

y	Observed outcome
$\bar{y}$	Mean observed outcome
μ	Mean outcome
s_h^2	Observed usual or naïve variance for study h
σ_h^2	Usual or naïve variance for study h
σ_b^2	Between-cluster variance
σ_w^2	Within-cluster variance
σ_t^2	Total variance

SSE	Residual or error sums of squares
SSB	Sums of squares between clusters
SSW	Sums of squares within clusters
SST	Total sums of squares
MSB	Mean squares between clusters
MSW	Mean squares within clusters
MST	Total mean squares
df	Degrees of freedom
w	Weight
Q	Q-statistic
η	Between-study variation in the precision of study estimates
φ	Non-centrality parameter for a t distribution
ψ	Test statistic

Multilevel Models (Chapters 2, 6 & 7)

y_l	Continuous outcome recorded at the patient level
y_l'	Transformed continuous outcome recorded at the patient level
α	Mean outcome for the l^{th} patient in the reference group
β	Matrix of coefficients for fixed baseline covariates
W_p	Coefficients for a fixed intervention-by-study covariate interaction
θ	Coefficient for the intervention effect
x_l	Matrix of indicator variables for fixed baseline covariates
s_{pl}	Indicator variables for a fixed study-level covariate
t_l	Indicator variable for the intervention
X_l	Indicator variable for studies with fully nested designs
C_l	Indicator variable for studies with clustering effects
T_l	Indicator variable for studies which standardised treatment provision
P_l	Indicator variable for studies which standardised patient characteristics
e_l	Random error for the l^{th} patient, also as $e_l^{(1)}$
ξ_l	Random error for the l^{th} patient in the intervention arm, also as $\xi_l^{(1)}$
e_{0l}	Random error for the l^{th} patient in the control arm, also as $e_{0l}^{(1)}$
e_{1l}	Random error for the l^{th} patient in the intervention arm, also as $e_{1l}^{(1)}$
$u_{therapist(l)}^{(2)}$	Random intercept for the l^{th} patient at the therapist level
$v_{therapist(l)}^{(2)}$	Random coefficient for the l^{th} patient at the therapist level
$u_{0therapist(l)}^{(2)}$	Random intercept for the l^{th} patient in the control arm
$u_{1therapist(l)}^{(2)}$	Random intercept for the l^{th} patient in the intervention arm
$q_{treat(l)}^{(2)}$	Random intercept for the l^{th} patient at the treatment level
$p_{therapist(l)}^{(3)}$	Random intercept for the l^{th} patient at the therapist level
$\tau_{study(l)}^{(3)}$	Random coefficient for the l^{th} patient at the study level, also as $\tau_{study(l)}^{(2)}$
σ_e^2	Sampling variance of the random error for the l^{th} patient
σ_ξ^2	Sampling variance of the random error for the l^{th} intervention patient
σ_{e0}^2	Sampling variance relating to e_{0l}
σ_{e1}^2	Sampling variance relating to e_{1l}
σ_u^2	Between-therapist variance relating to the random-intercept $u_{therapist(l)}^{(2)}$
σ_v^2	Between-therapist variance relating to the random-coefficient $v_{therapist(l)}^{(2)}$
σ_{uv}	Covariance between $u_{therapist(l)}^{(2)}$ and $v_{therapist(l)}^{(2)}$
σ_{0u}^2	Between-therapist variance relating to the random-intercept $u_{0therapist(l)}^{(2)}$
σ_{1u}^2	Between-therapist variance relating to the random-intercept $u_{1therapist(l)}^{(2)}$
σ_{01u}	Covariance between $u_{0therapist(l)}^{(2)}$ and $u_{1therapist(l)}^{(2)}$

σ_q^2	Within-therapist between-treatment variance relating to $q_{treat(l)}^{(2)}$
σ_p^2	Between-therapist variance relating to the random-intercept $p_{therapist(l)}^{(3)}$
τ^2	Between-study variance relating to the random-coefficient $\tau_{study(l)}^{(3)}$
ρ_u	Intraclass correlation coefficient relating to $u_{therapist(l)}^{(2)}$
ρ_0	Intraclass correlation coefficient for the control arm
ρ_1	Intraclass correlation coefficient for the intervention arm
ρ_q	Intraclass correlation coefficient relating to $q_{treat(l)}^{(2)}$

Meta-Analyses of ICC Estimates (Chapter 5)

$\hat{\rho}_h$	Intraclass correlation estimate observed within study h
ρ_h	Population intraclass correlation relating to study h
ρ	Mean population intraclass correlation
$\hat{\gamma}_h$	Blitstein *et al* transformed observed intraclass correlation
γ_h	Blitstein *et al* transformed population intraclass correlation
γ	Blitstein *et al* mean transformed population intraclass correlation
ε_h	Random error of ρ_h from ρ or γ_h from γ
e_h	Random error sampling of $\hat{\rho}_h$ from ρ_h or $\hat{\gamma}_h$ from γ_h
$\hat{\rho}_{A,h}$	ANOVA estimate of the study intraclass correlation
$\hat{\rho}_{ML,h}$	ML estimate of the study intraclass correlation
$\hat{\rho}_{A,h}^*$	Method-corrected ANOVA estimate of the study intraclass correlation
$\hat{z}_{F,h}$	Fisher's classical transformed estimate of the intraclass correlation
$\hat{z}_{A,h}$	Fisher's transformed ANOVA estimate of the intraclass correlation
$\hat{z}_{ML,h}$	Transformed ML estimate of the intraclass correlation
$\hat{z}_{KG,h}$	Konishi-Gupta transformed ML estimate of the intraclass correlation
$\tau_{\{\varepsilon_h\}}^2$	Between-study variance (i.e. variance of the study-level random errors)
$\sigma_{\{e_h\}}^2$	Within-study variance (i.e. variance of the within-study sampling error)
$T_{\{\hat{\rho}_h\}}^2$	Total variance relating to the raw estimate $\hat{\rho}_h$
$T_{\{\hat{\gamma}_h\}}^2$	Total variance relating to the transformed estimate $\hat{\gamma}_h$
$\sigma_{\{\hat{z}_{A,h}\}}^2$	Within-study variance relating to the transformed estimate $\hat{z}_{A,h}$
$\sigma_{\{\hat{z}_{KG,h}\}}^2$	Within-study variance relating to the transformed estimate $\hat{z}_{KG,h}$

Aggregate-Data Meta-Analyses of Intervention Effects (Chapters 6 & 7)

$\hat{\theta}_h$	Intervention effect estimate observed within study h
θ_h	Population intervention effect relating to study h
θ	Mean population intervention effect
ε_h	Random error of θ_h from θ

e_h	Random sampling error of $\hat{\theta}_h$ from θ_h
$\tau^2_{\{\theta_h\}}$	Between-study variance relating to the parameter θ_h
$\sigma^2_{\{\hat{\theta}_h\}}$	Within-study variance relating to the estimate $\hat{\theta}_h$
$T^2_{\{\hat{\theta}_h\}}$	Total variance relating to the estimate $\hat{\theta}_h$
$\hat{\theta}_w$	UMVUE of the pooled intervention effect
$\sigma_{\{\hat{\theta}_w\}}$	Standard error of the UMVUE of the pooled intervention effect
$\theta_{MD,h}$	Population mean difference relating to study h
$\sigma^2_{\{\hat{\theta}_{MD,h}\}}$	Sampling variance of the mean difference
$\theta_{SMD,h}$	Population standardised mean difference relating to study h
$\hat{\theta}_{Cohen's\,d,h}$	Cohen's *d* estimate of $\theta_{SMD,h}$
$\hat{\theta}_{Hedges'\,g,h}$	Hedges' *g* estimate of $\theta_{SMD,h}$
$\sigma^2_{\{\hat{\theta}_{Hedges'\,g,h}\}}$	Sampling variance of Hedges' *g*
$\sigma_{den,h}$	Standardising metric for the population standardised mean difference
a_{hi}	Known variance inflation factor determined by choice of metric
b_h	Ratio of the expectation of the sample metric to the population metric
$c(df)$	Hedges' small-sample bias correction
$Var(G_h)$	Ratio of the variance of the mean difference to the standardising metric
θ_{pth}	Standardised mean difference based on the pooled total SD
$\theta_{pth\mid\rho_{h0}=0}$	Pooled total standardised mean difference for partially nested studies
θ_{ph}	Standardised mean difference based on the pooled naïve SD
$s_{den,h}$	Standardising metric for the observed standardised mean difference
s^2_{pth}	Observed pooled total variance
s^2_{ph}	Observed pooled naïve variance
$g_{un,h}$	General version of $\hat{\theta}_{Cohen's\,d,h}$
$g_{adj,h}$	General version of $\hat{\theta}_{Hedges'\,g,h}$
$\sigma^2_{\{g_{un,h}\}}$	Sampling variance of $g_{un,h}$
$\sigma^2_{\{g_{adj,h}\}}$	Sampling variance of $g_{adj,h}$

Abstract

Randomised trials of complex interventions are typically designed, conducted, and analysed as if they are drug trials. Although there are many parallels there are also a number of important distinctions, which are seldom considered when designing individual trials. One of these concerns the involvement of therapists in delivering psychotherapy. Systematic reviews and meta-analyses provide an opportunity for exploring the full range and complexity of issues encountered in realistically complex situations. The first objective of the thesis was therefore to develop a conceptual framework for understanding the role of the therapist in trial designs. It was addressed by a review of the psychotherapy and statistical literatures structured according to the broad concepts of precision, internal and external validity and refined on the basis of a systematic methodological review of Cochrane reviews meta-analysing trials involving psychotherapy. The second objective was then to review, adapt, illustrate and compare methods for meta-analysing psychotherapy trials with nested designs. Methods for meta-analysing ICC estimates, absolute and standardised mean differences were adapted to allow for heteroscedasticity between treatments at the therapist- and patient-levels. These were illustrated using the example of counselling in primary care, with comparisons being made between aggregate and one-step approaches to the meta-analysis of individual-patient-data.

It was argued that the therapist has two roles in randomised trials. Firstly, they are one component of a multi-component intervention, and are thus a potential treatment variable. Second, the nesting of patients within therapists creates an additional level in the design, so the therapist is also an experimental unit. The inability to conceal or randomise allocations leads to observational components within the trial design and to heteroscedasticity which deserves more attention. Characterising complex interventions, like psychotherapy, with more than one treatment variable could facilitate greater understanding of their components, how they interact, which are important, to what extent, and for whom. It also brings what is currently referred to as *process research* into the remit of trials, enabling a more complete evaluation of the causal effects. The broad concept of multiple experimental units makes cluster-randomised, longitudinal, multi-centre, crossover, therapist- and group-based intervention trials special cases of a more general class of *multilevel trial*. All involve clustering effects; their nature and the appropriate statistical model varying according to the design. Methods were proposed for the meta-analysis of continuous outcome data for two-level nested designs. A general approach was adopted, where possible, to incorporate methods covering cluster-randomised trials and the Behrens-Fisher problem. It was clear that this is a relatively untouched methodological area in need of further exploration. For the same reasons as it became necessary to summarise clinical research, it is recommended that systematic methodological reviews be carried out on a larger scale in future.

Therapist Variation within Meta-Analyses of Psychotherapy Trials 17/12/2010
University of Manchester; Doctor of Philosophy; Rebecca Elizabeth Ann Walwyn

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Preface

I have previously been awarded a first for the BSc (Hons) in Psychology (1996-1999) at the University of Kent at Canterbury and a distinction for the MSc (Hons) in Social Statistics (1999-2001) at the University of Southampton. The first year of the MSc was a postgraduate certificate in statistics. My experience of statistical and methodological research consists of the work contributing to this thesis. I have 6 years of experience working as an applied statistician, 5 of which were gained prior to my being awarded an MRC fellowship to undertake this research. From October 2001, I was employed by the Sainsbury Centre for Mental Health, and located in the Biostatistics and Computing Department at the Institute of Psychiatry two days a week. In January 2003, I moved to the Mental Health and Neuroscience Clinical Trials Unit (MH&N CTU) at the Institute of Psychiatry, under the directorship of Professor Simon Wessely, as their first full-time statistician. In September 2006, I started a four year Medical Research Council Special Training Fellowship in Health Services and Health of the Public (ref G0501886). Simon Wessely was the sponsor and head of department for my fellowship, Dr Chris Roberts and Professor Graham Dunn were supervisors, and Professor Michael Dewey, Dr Tony Johnson, Dr Rachel Churchill, Professor Peter White, Professor Michael Sharpe, and Professor Trudie Chalder were named collaborators. As part of this fellowship, my PhD was externally registered with the University of Manchester, Chris Roberts and Graham Dunn as supervisors and Dr Roseanne McNamee as advisor. I was seconded to the MRC CTU, and to Professor Andrew Nunn and Tony Johnson, from March 2009 as part of the general training element of my fellowship.

Acknowledgements

My particular thanks go to Simon Wessely for making so much seem possible to me, to the Medical Research Council for making this possible, and to Chris Roberts for helping me to make this happen. Simon has given me much of my motivation for continuing to work in mental health, the MRC and Chris many of the tools with which to do so. I am greatly indebted to all three.

In addition, I would like to thank Graham Dunn, Roseanne McNamee, Michael Dewey, Andrew Nunn, Tony Johnson, Rachel Churchill, Trudie Chalder, Michael Sharpe, Peter White and Jonathan Cook for their encouragement and support; Rosemary Bailey and Terry Speed for helpful discussions at a workshop on experimental design hosted by the Newton Institute of Mathematical Sciences in July 2008; Julian Higgins and Scott Baldwin for sharing papers they had in preparation; Pamela Gillies, Clair Chilvers, Karin Friedli, Ian Harvey, Adrian Hemmings, Michael King, Peter Bower, Roslyn Corney and Sharon Simpson for sharing their trial datasets; and finally Simon Thompson (MRC BSU, Cambridge) and Steve Roberts (Manchester) for their comments while examining this thesis.

Chapter 2 has previously been published in a similar form in: Walwyn, R., Roberts, C. Therapist variation in randomised trials of psychotherapy: Implications for precision, internal and external validity. *Statistical Methods for Medical Research*, 2010, ***19***, 291-315. Trudie Chalder (TC), Chris Roberts (CR) & Michael Dewey (MD) reviewed sections of Chapter 3. CR wrote the first draft of Section 2.3 for this paper.

1 INTRODUCTION

The randomised controlled trial (RCT) is usually considered the gold standard research design for evaluating the causal effects of therapeutic interventions. The methodology used has largely been developed in the context of simple interventions, such as drug treatments. Randomised trials of complex interventions are, thus, typically designed, conducted and analysed as if they were drug trials. Although there are many parallels there are also a number of distinctions. These have been explored in the recently updated guidance on the development and evaluation of complex interventions[1, 2], in which complex interventions are defined as containing several interacting components. One of the distinctions identified concerned the involvement of care providers in the intervention delivery. In psychotherapy research, therapist involvement is particularly central to the design. The nature and implications of this are the topic of this thesis.

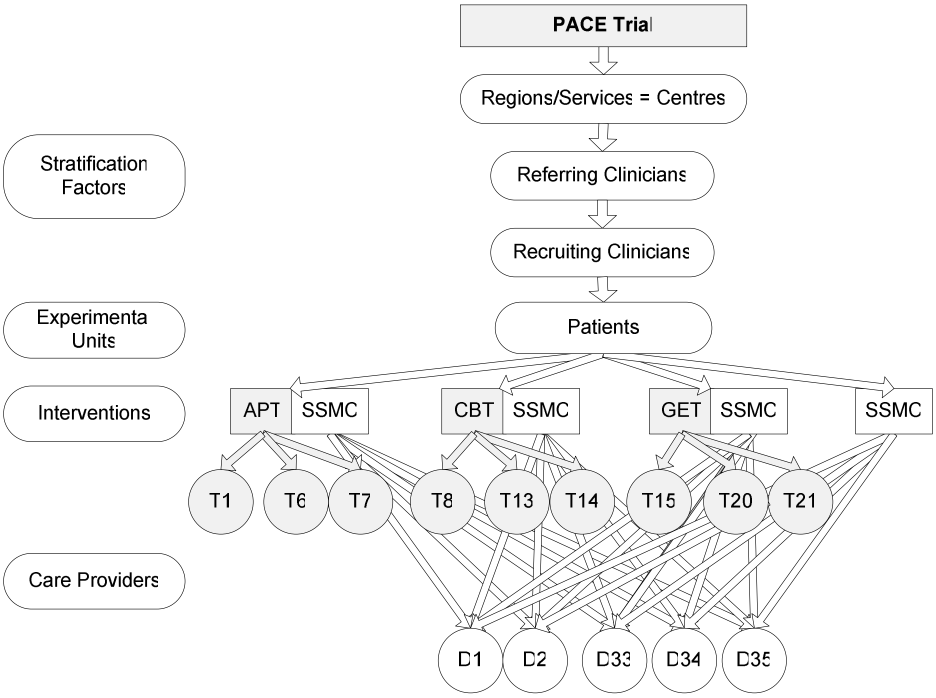
1.1 Motivating Example: The PACE Trial

Much of the motivation for this thesis originated from discussions, formal and informal, regarding care providers in the PACE trial[3]. The PACE trial (Pacing, graded Activity, and Cognitive behaviour therapy: a randomised Evaluation) is near completion having randomised 641 patients with a clinical diagnosis of Chronic Fatigue Syndrome/Myalgic Encephalomyelitis or Encephalopathy (CFS/ME) to one of four interventions:

(i) Adaptive Pacing Therapy plus Standardised Specialist Medical Care (APT);
(ii) Cognitive Behavioural Therapy plus Standardised Specialist Medical Care (CBT);
(iii) Graded Exercise Therapy plus Standardised Specialist Medical Care (GET); or
(iv) Standardised Specialist Medical Care (SSMC)

Patients were recruited from six secondary care services within the UK between March 2005 and November 2008. All consecutive patients with a clinical diagnosis of CFS/ME referred to the services during this period were screened for eligibility. Referrals were made by clinicians outside the services, including general practitioners in primary care. Screening was carried out by clinicians within the services, and "centres" were defined by the clinical service. Either clinician could have been a "centre", instead or as well. Figure 1.1 gives a schematic illustration of the relationship between treatments and care providers in the PACE trial.

Figure 1.1 Schematic Diagram of the PACE Trial Design



Note: T1 to T21 are therapists, D1 to D35 are doctors

Allocation of interventions to patients was by minimisation with a random component[4], stratified by centre, and whether patients met CDC criteria for CFS, London criteria for ME, or had a current depressive disorder, respectively. Medical care (SSMC) across all four intervention arms was provided by centre-specific doctors, including psychiatrists, psychiatric registrars and physicians: none were research doctors employed specifically for the trial. Therapy (APT, CBT and GET) was delivered by centre-specific therapists. There was a debate at the planning stage about whether therapists should provide all or only one of the therapies. It was thought that contamination between the therapies should be minimised and that this would be more easily achieved if different therapists provided each therapy. Besides this, it was believed that each therapy might naturally be championed by a different professional group: APT by occupational therapists, CBT by psychotherapists, and GET by physiotherapists. Hence, by design, one therapist per centre was specifically employed to provide each therapy, so allocation of therapists to patients was determined by the centre and allocated intervention. SSMC doctors were allocated to patients using the usual practice within each service.

The therapies (APT, CBT and GET) were intended to be structurally equivalent, in that the number of initial training days for the therapists and the number and length of the therapy sessions were to be roughly constant. Therapy provision was standardised to minimise variability between therapists within therapies. As equipoise was promoted to all four interventions throughout the trial, medical care was also standardised. Since the service was a stratification factor at randomisation, it was to be included as a fixed effect in the primary analysis. Clearly, this trial shares many features with drug trials.

1.2 Focus on Psychotherapy

The involvement of people in the provision of complex interventions is not restricted to psychotherapy or even to medicine. The issue generalises to education, physiotherapy, surgery, occupational therapy, complementary therapies and beyond. However, there are potentially important differences in the research questions and practices that could affect the way in which this issue is addressed. Each discipline has also developed its own methodological literature, published largely in subject-specific journals, separated from the statistical and mainstream trial methodology literature. A trade-off was made, as a consequence, between breadth and depth when defining the scope of this thesis. To enable the approach to be broad and inclusive, and to use experience of the field, the application area was limited to psychotherapy. By 1997, over 450 distinct forms of psychotherapy[5] and 100 distinct forms of group psychotherapy[6] had been identified. The thesis did not focus on any specific modality.

Defining what is meant by the term *psychotherapy* is far from straightforward, in large part because there is no universally accepted definition to draw on. Broadly speaking, psychotherapy is a complex therapeutic intervention based on psychological principles. Wampold[7] adds to this, that psychotherapy involves primarily face-to-face interactions between a trained therapist and a client and is typically individualised to the client and their disorder, problem or complaint, but this is perhaps too restrictive. Psychological interventions delivered over the phone or via a computer were thus included, as well as those delivered by paraprofessionals, such as General Practitioners (GPs) or nurses. Self-help interventions delivered with the guidance of a therapist were also included, although purely self-help interventions, such as bibliotherapy, were not. Interventions involving the clients in groups, or multiple therapists per client, were included, as were standardised or structured therapies. Psychotherapy is commonly grouped according to its format and theoretical model. Individual psychotherapies are often distinguished from group psychotherapies as different formats. Categories based on the theoretical orientation include psychotherapies using psychodynamic, cognitive and behavioural, systemic, humanistic, supportive, integrative or eclectic principles.

1.3 Conceptualising the Role of Therapists in Trial Designs

Perhaps the main reason why therapists are overlooked to a large extent in the design and analysis of psychotherapy trials is the widespread use of the drug metaphor within the clinical trial literature. The procedures used by the therapist are equated with the drug whilst the therapist is equated with the prescribing doctor. The reasoning behind some of the features of drug trials is implicit, so the consequences of substituting or omitting these features can be unclear. They are also seldom questioned during the design phase of individual psychotherapy trials. Legitimate concerns about this have led some to question how appropriate randomised controlled trials are for evaluating psychotherapy. It is possibly more accurate to infer that such questions apply to the trial design rather than to randomisation *per se*. The first objective of this thesis was, therefore, to develop a conceptual framework for understanding the role of therapists in psychotherapy trial designs.

1.3.1 Stratification versus Treatment Factors

One source of confusion relates to the distinction between *stratification* and *treatment factors*. Any therapist variability remaining after standardisation might be regarded, for instance, as subsumed within centre variability, so attention is focused on the centres, making reference to the literature on centre effects[8-19]. Indeed there may be a one-to-one relationship between centres and therapists. Relationships between the patients, treatments and centres may differ from those between the patients, treatments and therapists however. Moreover, centres are allocated to patients *prior* to randomisation, while the therapists are assigned to the patients before treatment but generally *after* randomisation. On this basis, centres can be regarded a potential stratification factor and the therapists a potential treatment factor.

In the PACE trial, the referring and the recruiting clinicians were potential stratification factors. The SSMC doctors and the therapists were potential treatment factors. Where the recruiting clinician was also the SSMC doctor, the line between these two types of factor might, at first, seem blurred. It is not, however, because the SSMC doctors only provided medical care, and did so for all four arms of the trial, so medical care is a co-intervention given by a separate sample of care providers. This division makes the role of the SSMC doctors akin to that of the care providers in a drug trial. If the recruiting clinicians had provided both the therapy and medical care, they would have remained a potential stratification factor because they were non-randomly assigned to patients before treatment allocation. If the SSMC doctors had provided both the medical care and the therapy, they would have remained a potential treatment factor as they were assigned to the patients alongside the interventions. Since their assignment might be random or non-random depending on the choice of design, they are regarded here as "potential" treatment factors. The focus of the thesis is on care providers as potential treatment factors. Issues relating to care providers as potential stratification factors are similar to those for multi-centre trials.

1.3.2 Performance Bias, Common Factors and Non-Specific Effects

In drug trials, the research question usually relates to the drug rather than to the drug as a function of the context in which it was provided. This is because such trials aim to make generalisations beyond their context. The medical care given by the doctor, their characteristics, and the nature of the doctor-patient relationship are not of interest for this research, neither are the clinical service or the geographical location. The physical separation between the drug and its context permits its identity to be blinded to everyone involved in the trial including patients, care providers and outcome assessors. This is achieved via a placebo, if the comparison is no drug, or by making the appearance of the drugs identical otherwise. The consequence is that it is then reasonable to make the assumption that there is no interaction between the treatment effect and the care providers when analysing patient outcomes. This in turn simplifies the statistical model but it has implications for sample size and interpretation. The presence of systematic differences in the provision of the intervention or any co-interventions, or performance bias, questions the validity of this assumption and signifies a more complex underlying model incorporating this interaction. The success of blinding is, therefore, a function of the size of this interaction. Since a drug trial is not powered to detect it, knowledge of how the trial was conducted is usually used to judge its size, and the credibility of the trial results.

It is not possible to blind a therapist to the procedures they use in psychotherapy. This makes the role of the therapist in a psychotherapy trial fundamentally different to that of a prescribing doctor in a drug trial, in that it is more accurate to equate a therapist plus the procedures with a drug, as therapists are part of the treatment rather than its context. It also questions the relevance of the concept of performance bias in this setting because the assumption that there is no interaction between the procedures and the care provider is simply untenable, even at the design stage. However, instead of just accepting this, the field has responded by creating placebo psychotherapies to control for the placebo effect[20, 21]. As is to be expected, this has caused considerable debate. Since the reason for a placebo no longer applies, one aspect of this debate concerns the definition of a placebo in this setting (see Horvath[22]). A variety of terms has been used, reflecting the lack of consensus, including "non-specific" control and "common factors" control. A recent issue of *Clinical Psychology-Science & Practice* was devoted to discussing the rationale behind them[23-27]. It is clear from this that the issue the field is attempting to address is not how to define a psychotherapy placebo, but how to evaluate the individual causal effects of a *multi-component* or *multivariate treatment*. As drugs are single-component treatments, included as a single variable in a statistical model, the drug metaphor falls down. That is, unless, perhaps, you extend the analogy to include combinations of drugs in an open-label trial, where each drug is entered as a separate variable in the model, along with relevant interaction effects.

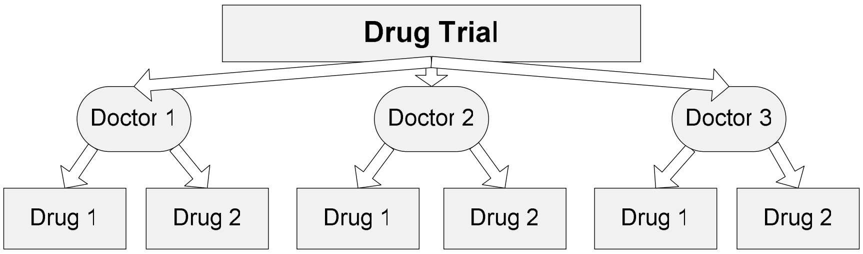
A second aspect of this debate regards the relative importance of the so-called specific and common factors to the total effect of psychotherapy. This has become an emotive issue because specific factors tend to be equated with drugs which are independent of the therapist, and common factors with their context, including the therapist. Luborsky *et al*[28] first asserted that the total effects of most, if not all, psychotherapies do not differ, referring to the Dodo bird's verdict in Alice in Wonderland that "Everyone has won, and all must have prizes". This has been used as a vehicle to raise the profile of common factors, one of which is thought to be the therapist[7, 29-36]. This is valuable, but it can be unhelpful, when done at the expense of the procedures that generally come under the specific factors. Kazdin[26], instead, posed the following challenge to the field:

> "Let us do experimental manipulations to explore what can be done to improve therapy in relation to both specific and common factors. In the process, we shall learn much about the mechanisms of change and how to better help patients." (p.186)

1.3.3 Multilevel Trial Designs

Blinding of the prescribing doctors in a drug trial generally means they provide all the drugs in the trial so that the treatment effect is estimated within each doctor, although the estimation procedure does not generally take account of this (see Figure 1.2). If there is no interaction between a drug and the doctor, the population treatment effect will be identical for all doctors. Any variability that is observed in the treatment effects across doctors will arise due to chance alone. If the doctors are included in the model, either as a fixed or a random stratification effect, this will increase the precision of the treatment effect estimate where the outcome varies between the doctors enough to warrant the loss of residual degrees of freedom.

Figure 1.2 Simplified Schematic Diagram of a Drug Trial



The nesting of patient outcomes within geographical regions, services, recruiters, care providers, outcome assessors, informants, and patients creates additional levels within the designs of all trials, not just those that employ cluster-randomisation. Interactions are possible between the treatment effect and each cluster type, leading to treatment-related clustering effects. A blinded drug trial, that has concealed randomisation, uses blinding of the assignment, treatment, and outcome assessment to rule out possible sources of treatment-related clustering, and any higher-order interactions, leaving the clusters stratifying the treatment effect. As with fixed stratification factors, a decision could be made to account for the clusters in the design and subsequent analysis based on the extent to which doing so is expected to increase the precision of the treatment effect estimate.

If blinding cannot be used to remove important sources of treatment-related clustering then additional concealed randomisations could be considered instead. When they are not, the treatment effect is conditional on the populations from which the clusters are sampled. Generalisations are hence restricted to the treatment context, but they need not be further restricted to the trial context. This means that the process of ensuring the trial design adequately addresses the research question is more complicated, but it does not mean that trials of complex interventions are either inherently more biased or less robust or credible than drug trials. As the therapist is arguably the most important source of treatment-related clustering in psychotherapy trials, therapists were focused upon in this thesis.

1.4 Incorporating Therapist Variation into Meta-Analyses

Once it is accepted that treatment-related clustering associated with the care providers is neither entirely avoidable or should necessarily be avoided, the next challenge is the shared lack of experience within the medical statistics and psychotherapy fields of fully taking the therapist into consideration when designing trials. One of the first steps that should be taken when planning trials is to systematically review the existing evidence-base. This may include an assessment of the quality of previous trials, pooling of data across trials, or both activities. This presents an opportunity, as one way of raising the general level of experience and of establishing the range and complexity of issues that may be encountered, is for methodologists to liaise more closely with those conducting systematic reviews. Meta-analyses also provide a useful parallel for analyses of multi-centre trials, as studies in a meta-analysis corresponds to the centres in a multi-centre trial. Early-phase psychotherapy trials, of necessity, are small, involve few therapists, and can completely confound therapists with centres. Due to their size, meta-analyses provide a means of gaining experience of fitting realistically complex statistical models, applicable to trials such as PACE, with real datasets. The second objective of the thesis was consequently to review, adapt, illustrate and compare methods for meta-analysing trials involving psychotherapy to take account of the therapist.

1.5 Thesis Overview

In this chapter the PACE trial has been used to motivate a discussion of the limitations of the drug metaphor for randomised trials of complex interventions, focusing interest on the therapist. The scope of the thesis was outlined, limiting the application area to psychotherapy, with two objectives being defined. The first was to provide a complete conceptualisation of the role of therapist in psychotherapy trial designs. A distinction was made here between stratification and treatment factors, with the therapist defined as potential treatment factor. Performance bias was defined as an interaction between the therapist and the intervention effect, implying the presence of multiple interacting treatment factors in trials of complex interventions. Nesting of patient outcomes within therapists was argued to create multiple experimental units in psychotherapy trials and lead to treatment-related clustering. The second objective was to explore methods for meta-analysing psychotherapy trials incorporating therapists as a factor.

Chapter 2 expands the conceptual framework using the familiar concepts of precision, internal and external validity, reviewing the psychotherapy and statistical literatures. Research questions that relate to therapeutic approaches, therapist characteristics and packages of the two are unified. Three basic trial designs are introduced which map to the relationship between interventions and therapists. Implications for the precision of treatment effects are given for statistical analyses and sample size. More complex trial designs are then described for trials with multiple therapists per patient. Implications for internal and external validity are outlined and conclusions are drawn.

The framework is used in Chapter 3 to structure a systematic methodological review of Cochrane reviews of comparative studies involving psychotherapy. The frequency with which specific trial designs appear is explored, together with the range and complexity of issues that would be encountered by meta-analysts in a psychotherapy setting. The search strategy and eligibility criteria for selecting the systematic reviews are provided. The systematic reviews meeting the criteria and the studies within those reviews are described. Data is extracted and summarised in support of reviewers' recognition of the concept of therapist variation, and also their recognition of the precision, internal and external validity implications of therapist variation.

Chapter 4 uses the methodological review as a sampling frame to select counselling in primary care as an example to illustrate the methods described in the remainder of the thesis. The Cochrane review[37] is introduced along with the example meta-analysis, the trials included, and the individual-patient-data (IPD). This consists of a background to the field of counselling in primary care, a summary of the review methodology, its original analysis and published results. The trial methodology is outlined on a trial-by-trial basis, providing the published summary data and analyses. The practicalities of obtaining the individual-patient-data relating to these trials are then described.

Chapters 5, 6 and 7 address the second objective of reviewing, adapting, illustrating and comparing methods of meta-analysis. Chapter 5 considers methods for obtaining and pooling intraclass correlation coefficients (ICC), adapting a random-effects meta-analytic approach proposed by Blitstein *et al*[38] for cluster-randomised trials. The lack of relevant estimates is highlighted and the small-sample issues introduced. *Adhoc* versus systematic strategies for obtaining ICC estimates are then contrasted. Blitstein *et al*'s[38] method is described and adapted to take account of biases arising from the method of estimation, the skewed sampling distribution and bounding negative estimates, and of variability in cluster sizes within studies. Relevant external estimates are obtained and compared to the internal study estimates. The steps required to pool these estimates are then described, comparing existing approaches to those proposed in the chapter. The chapter ends with a discussion of the availability of study estimates, the need for further work, the role of sampling variation and the implications for study reporting.

Chapter 6 compares aggregate-data and one-step multilevel approaches to the meta-analysis of absolute mean differences, adapting the methods suggested by Kwong and Higgins[39] and Sidik and Jonkman[40]. The rationale is given for allowing for treatment-related clustering effects, and for taking account of any between-study heterogeneity. Standard fixed and random effects meta-analysis models are presented and extended, relaxing independence and homoscedasticity assumptions and allowing for imprecision in the estimated weights and the use of finite samples. Equivalent one-step models are given, including those exploring predictors of between-study heterogeneity in the fixed or random effects. These are illustrated and compared with a subset of the counselling in primary care trials. The programming code is given for the one-step models in an appendix.

Chapter 7 compares aggregate-data and one-step multilevel approaches to the meta-analysis of standardised mean differences, adapting those proposed by Huynh[41], White and Thomas[42], Hedges[43] and Goldstein *et al*[44]. Issues concerning the choice of metric are introduced for different model assumptions and the need for a general approach is discussed. Standard fixed and random effects meta-analysis models are first presented and a general approach is developed that allows for clustering and heteroscedasticity. Equivalent one-step models are then provided and the preparation of the data needed for these models described. These are illustrated and compared using all the trials of counselling in primary care. Programming code is again given for the one-step models in an appendix.

A discussion of the thesis is found in Chapter 8. This is structured according to the two objectives, placing the thesis in the context of the existing literature. Areas for further research are identified.

2 THERAPIST VARIATION WITHIN RANDOMISED TRIALS OF PSYCHOTHERAPY: IMPLICATIONS FOR PRECISION, INTERNAL AND EXTERNAL VALIDITY

2.1 Introduction

Psychotherapy is defined as a non-pharmacological intervention delivered by therapists based on psychological principles. The intervention of interest in a psychotherapy trial broadly lies "somewhere in the therapist and his behavior"[45] (p.128). While it is generally reasonable to assume a drug is manufactured uniformly, this assumption is less plausible for psychotherapy both generally and in the context of randomised trials. Variation is expected in psychotherapy content and format across patients[45-47], across time within patients[45] and across time within therapists[48, 49]. Akin to doctors prescribing drugs, therapists are able to influence patient adherence to psychotherapy and the provision of co-interventions. Therapists, however, are additionally able to influence the content of psychotherapy via their skill, expertise, competence or fidelity to the therapy model and its format via their personal characteristics.

The notion that patient outcomes vary between therapists has been recognised by psychotherapy researchers and clinicians since the origin of the field[7], although not universally as Kiesler[45] highlighted. Methods for studying the contribution of therapists to patient outcomes have changed over time[50]. Up to 1960, research was mainly qualitative involving the elicitation of expert opinions on therapist characteristics deemed to be important (e.g. Luborsky[51]). Since then associations between therapist characteristics and patient outcomes have been studied (see reviews by Parloff *et al*[52] and Beutler *et al*[53]) and comparisons made of patient outcomes between individual therapists (see Ricks[54], Howard *et al*[55], Orlinsky & Howard[56], Brooker & Wiggins[57], Luborsky *et al*[58, 59], McLellan *et al*[60] and Shapiro *et al*[61] for early examples; Okiishi *et al*[62, 63], Wampold & Brown[64], McKay *et al*[65], Lutz *et al*[66], Baldwin *et al*[67] and Dinger *et al*[68] for more recent examples; and see the March 2006 edition of *Psychotherapy Research*[69-74] and re-joiners[75-77] for a discussion of the use of multilevel models in this context).

Despite awareness of therapist variability, the statistical and wider conceptual implications of therapist variation for psychotherapy trials have not been widely recognised. Research on the relationship between individual therapists and patient outcomes has also been separated from randomised trials[78]. While psychotherapy researchers readily recognise that average patient outcomes may vary between therapists, they infrequently appreciate that this is equivalent to patient outcomes being *clustered* within individual therapists leading to intra-therapist correlation. This violates the assumption of statistical independence in much the same way as nesting of individuals within clusters does in cluster randomised trials. Ignoring clustering of patients within therapists tends to lead to over-precise estimates of the intervention effect and an increase in the type I error.

The clustering implications of therapist variability were outlined firstly within the psychotherapy literature by Martindale[79] and then by Crits-Christoph and Mintz[80]. Subsequently Roberts[81], Lee and Thompson[82, 83] and Roberts and Roberts[84] have brought the issue to the attention of the mainstream medical statistics community. This widening awareness has culminated in the inclusion of items specifically relating to therapist or *care provider* variation in the extended CONSORT guidelines for the reporting of non-pharmacological trials[85]. The publication of these guidelines should motivate increased interest in this topic both in psychotherapy trials and more generally.

The potential impact of therapist variation on the design, analysis and reporting of randomised trials is threefold. While attention has focused primarily on the implications for the precision of treatment effect estimates, therapist variation also has implications for internal and external validity. This chapter uses all three aspects as a broad framework for understanding the implications of therapist variability, drawing together and building upon the associated psychotherapy and statistical literatures. In so doing, parallels with other trial designs and associated methods of analysis are made.

2.1.1 Research Questions

Therapists and their behaviours may be viewed as potentially interacting components of a "complex intervention"[86]. Based on the drug metaphor of treatment, trialists often emphasise particular therapist behaviours or therapeutic approaches, such that the therapist is considered part of the therapeutic context[7]. At other times, emphasis is placed on *packages* of particular therapeutic approaches and "the therapists who both choose to and are chosen to administer them"[48] (p.308). In which case, the therapist can be viewed as an integral part of the intervention itself[81].

Figure 2.1 illustrates three types of comparison between interventions:

1. Where the *therapeutic approach* is of interest, a comparison might be made between **A** and **B** or between **C** and **D**. So for example, counselling could be compared to advice, both provided by counsellors.
2. Where a *package* is of interest, a comparison might be made between **A** and **D** or between **B** and **C**. As an example, counselling provided by counsellors could be compared to advice provided by general practitioners.
3. Where particular *therapist characteristics* are of interest, a comparison might be made between **A** and **C** or between **B** and **D**. For example, counselling provided by counsellors and general practitioners could be compared.

The research question will determine the comparisons made, and these should in turn determine the trial design. In all three cases, variation may be anticipated in the provision of psychotherapy between individual therapists and therefore also potentially in patient outcomes across therapists. Whilst lack of clarity is a precursor for poor design, these comparisons may be of interest separately or in combination.

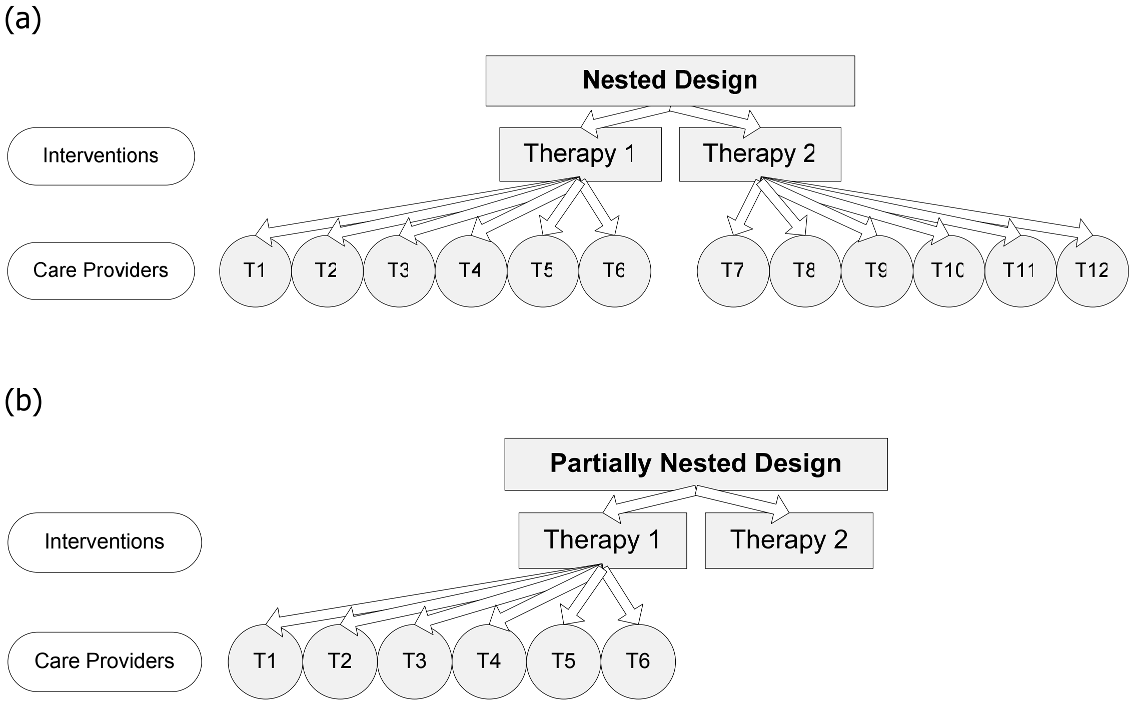
Figure 2.1 Example Comparisons

		Therapeutic Approaches	
		Counselling	Advice
Therapist Characteristics	Counsellors	A	B
	GPs	C	D

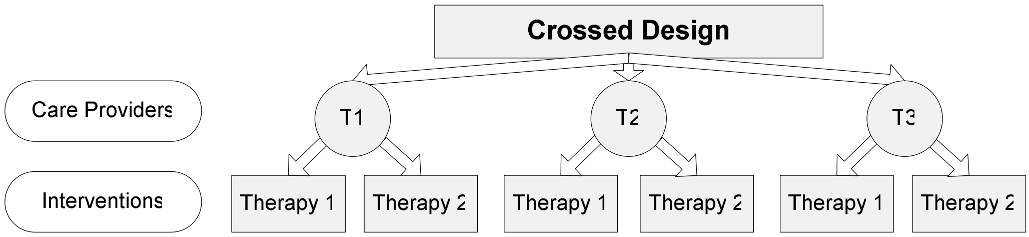
2.2 Trial Designs

The comparison of two interventions, where each patient receives psychotherapy from just one therapist, is considered first. In the *nested design* (Figure 2.2a) each therapist provides the intervention within just one intervention arm, a design described as *hierarchical* by Martindale[79]. As an example, Schnurr *et al*[87] compared Prolonged Exposure to Present-Centred Therapy for women with Posttraumatic Stress Disorder where each therapeutic approach was provided by a different sample of therapists. In the special case where there is no therapist involvement in one of the arms[81, 83, 84] the design can be described as *partially nested* (Figure 2.2b). This may arise where the control intervention is no treatment, a waiting list or self-help. For example, Kubany *et al*[88] compared Cognitive Trauma Therapy with a waitlist control for women with Posttraumatic Stress Disorder. Alternatively, in the *crossed design*[79] each therapist provides psychotherapy in both arms (Figure 2.2c). An example of this is the comparison of Cognitive Behavioural Therapy with Nondirective Supportive Therapy for sexually abused children provided by therapists experienced in treating this patient group[89].

Figure 2.2 Nested, Partially Nested and Crossed Designs



(c)



Note: T1 to TK are therapists

The crossed design is restricted to situations where therapists are able to deliver both interventions. It is really only applicable where different therapeutic approaches are compared, or where psychotherapy may be viewed as a co-intervention therefore. The latter arises where the combination of psychotherapy and a drug is compared to psychotherapy plus a placebo, for example, or where two formats of the same psychotherapy (e.g. telephone vs. face-to-face) are compared, as was the case in Lovell *et al*[90]. One advantage of separating the role of doctors and therapists in a trial assessing the efficacy of a psychotherapy-drug combination is that it may then be possible to blind therapists to the drug/placebo component of the intervention and therefore also to treatment status.

Parallels can be drawn between parallel-group/crossover designs and nested/crossed designs at the level of the therapist. Interventions are allocated to patients within the former but to therapists within the latter. In a parallel-group trial patients are nested within interventions, while it is the therapists who are nested within interventions in a nested design. In a crossover trial intervention sequences are allocated to patients. Similarly one may consider a sequence of intervention assignments as being allocated to therapists in a crossed design, so that patients within therapists correspond to periods within patients in a crossover trial. The point in the intervention sequence at which a patient is assigned to the therapist is therefore equivalent to the *period* in a crossover trial. If each therapist were to see just two patients, intervention sequences might be allocated to therapists much as they are to patients in an AB/BA crossover design. Possible carryover effects within therapists from the first patient to the second would need to be considered and differential carryover from one intervention to the other would invalidate the use of both designs. However, as therapists typically treat more than two patients, intervention sequences would generally be longer in a crossed design so that these designs are more likely to be comparable to replicate crossover trials[91]. In the special case where intervention sequences are formed of replicates of shorter sequences it may be possible to separate order effects from those over time within therapists.

Parallels can be drawn between completely randomised, matched-pair and stratified cluster randomised trials and different nested designs as well by equating the therapist in a nested design to the cluster within a cluster-randomised design. The allocation of interventions to clusters is blocked or stratified (e.g. by centre) in matched-pairs and stratified cluster randomised designs, with matched-pairs designs being a special case of the stratified design, where the number of clusters and interventions within strata is equal. Limitations of matched-pairs designs[92] therefore also apply to nested designs where the number of therapists equals the number of interventions within a centre. Likewise a cluster randomised crossover design[93, 94] corresponds to a crossed design. A crossed design is also related to matched-pairs cluster randomised designs, but now the therapist corresponds to a pair of clusters rather than the cluster, which is perhaps a more tenuous analogy. Other analogies for a crossed design are multicentre trials with the therapist corresponding to the centre, and meta-analyses with therapists corresponding to studies.

2.3 Implications for Precision

2.3.1 Statistical Analyses

Suppose y_l is a continuous outcome for the l^{th} patient within a typical drug trial. An analysis of covariance model can be written as:

$$y_l = \alpha + \theta t_l + \beta x_l + e_l \quad (2.1)$$

where t_l is an intervention indicator variable, θ is the intervention effect, e_l is $N(0, \sigma_e^2)$ the patient level error term, and x_l and β are matrices representing fixed patient or therapist level baseline covariates and their coefficients. For simplicity of presentation let α_l equal $\alpha + \beta x_l$. Consider now a nested design where m_j patients are treated by the j^{th} therapist. Using Goldstein's[95] notation, between-therapist variation can be represented by a random effect $u^{(2)}_{therapist(l)}$ with distribution $N\left(0, \sigma^2_u\right)$. The outcome for the l^{th} patient treated by the j^{th} therapist is then given by a random intercept model

$$y_l = \alpha_l + \theta t_l + u^{(2)}_{therapist(l)} + e^{(1)}_l \quad (2.2)$$

In this notation the bracketed superscript refers to the level of the random effect and $therapist(l)$ in the subscript is the mapping of patients to therapists. Between-therapist variation σ^2_u inflates the standard error of the intervention effect estimate $\hat{\theta}$ in the same way that between-cluster variation does in cluster randomised trials. Intra-therapist variability is measured by the intraclass correlation coefficient (ICC) ρ_u defined by $\rho_u = \sigma^2_u / \left(\sigma^2_u + \sigma^2_e\right)$. This model, suggested by a number of authors[49, 79-81, 96, 97], assumes the clustering effect is the same for both interventions.

Where therapist variation is considered a nuisance, a two-stage analysis has been proposed[79, 80, 97], using a test of non-zero therapist effect to determine whether to include a random effect for therapist. Model (2.2) is then only used if the null hypothesis is rejected. Elkin[78] indicated that such analyses may be carried out but not reported because they are seen as a "necessary, but preliminary, step for any outcome analyses" (p.13). Even with the suggested use of significance levels of 20% to 30% for the preliminary analysis[98, 99] this strategy has not been recommended by Lee and Thompson[82] or by Roberts and Roberts[84]. The preliminary test will have low power to detect intra-therapist correlation coefficients with potentially considerable design effects, a point also made by Donner and Klar[100] in the context of cluster randomised trials. What is more, in common with other pre-testing procedures, for example tests for carry-over effect in crossover trials, this analysis strategy misuses hypothesis testing, as failure to reject the null hypothesis of no effect does not justify its acceptance[84].

Suppose now that between-therapist variability differs across intervention arms. This could occur because one intervention is more suited to standardisation than the other, resulting in average patient outcomes being more homogeneous for one arm than the other. Lee and Thompson[83] suggest a random coefficient model

$$y_l = \alpha_l + \theta t_l + u^{(2)}_{therapist(l)} + v^{(2)}_{therapist(l)} t_l + e^{(1)}_l \quad (2.3)$$

where $v^{(2)}_{therapist(l)}$ is a $N\left(0, \sigma^2_v\right)$ for individually randomised trials with clustering effects. Random effects at the same level may be correlated, so $u^{(2)}_{therapist(l)}$ and $v^{(2)}_{therapist(l)}$ may have a covariance term, say σ_{uv}. In a nested design t_l does not vary within therapists so only one of the random effect parameters σ^2_v or σ_{uv} is identified[83]. In this situation, estimation procedures add constraints such as setting either σ^2_v or σ_{uv} to zero or setting σ_{uv} equal to $-\sigma^2_u$. Separate intraclass correlation coefficients for each intervention arm are then

$$\rho_0 = \sigma^2_u / \left(\sigma^2_u + \sigma^2_e\right) \text{ and } \rho_1 = \left(\sigma^2_u + 2\sigma_{uv} + \sigma^2_v\right) / \left(\sigma^2_u + 2\sigma_{uv} + \sigma^2_v + \sigma^2_e\right)$$

Where $\sigma_{uv} = 0$, $\sigma^2_u + \sigma^2_v$ is forced to be greater than or equal to σ^2_u, so it is more appropriate to set $\sigma^2_v = 0$ here as a covariance may be negative, while a variance cannot be.

An alternative parameterisation[81, 83] of model (2.3) is

$$y_l = \alpha_l + \theta t_l + u^{(2)}_{0therapist(l)}\left(1 - t_l\right) + u^{(2)}_{1therapist(l)} t_l + e^{(1)}_l \quad (2.4)$$

where $u^{(2)}_{0therapist(l)}$ and $u^{(2)}_{1therapist(l)}$, distributed $N\left(0, \sigma^2_{0u}\right)$ and $N\left(0, \sigma^2_{1u}\right)$ respectively, are now random intercepts for each intervention arm. Whilst this model is slightly more cumbersome to fit, it is easier to interpret where the intervention is a package or therapist characteristic. It also gives the between-therapist variance estimates directly for each intervention arm, which avoids constraints on their relative size and may be convenient for reporting. Since $u^{(2)}_{0therapist(l)}$ and $u^{(2)}_{1therapist(l)}$ relate to independent samples their covariance is zero, consistent with model (2.3). The intra-therapist correlation coefficients for each intervention arm are now simply $\rho_0 = \sigma^2_{0u} / \left(\sigma^2_{0u} + \sigma^2_e\right)$ and $\rho_1 = \sigma^2_{1u} / \left(\sigma^2_{1u} + \sigma^2_e\right)$. In a nested design the relationship between the parameters of models (2.3) and (2.4) is $\sigma^2_{0u} = \sigma^2_u$ and $\sigma^2_{1u} = \sigma^2_u + 2\sigma_{uv}$.

Models (2.1) to (2.4) assume a common patient level variance σ^2_e across intervention arms. Heteroscedasticity at the patient level may bias estimates of the between-therapist variance[84]. For this reason Roberts and Roberts[84] suggest the two-level heteroscedastic model

$$y_l = \alpha_l + \theta t_l + u^{(2)}_{therapist(l)} + v^{(2)}_{therapist(l)} t_l + e^{(1)}_l + \xi^{(1)}_l t_l \quad (2.5)$$

Where $\xi^{(1)}_l$ is $N\left(0, \sigma^2_{\xi}\right)$. A recommended parameterisation of a two-level heteroscedastic model for a nested design is, however,

$$y_l = \alpha_l + \theta t_l + u^{(2)}_{0therapist(l)}\left(1-t_l\right) + u^{(2)}_{1therapist(l)} t_l + e^{(1)}_{0l}\left(1-t_l\right) + e^{(1)}_{1l} t_l \quad (2.6)$$

Separate intra-therapist correlation coefficients for the arms are then

$$\rho_0 = \sigma^2_{u0} \big/ \left(\sigma^2_{u0} + \sigma^2_{e0}\right) \text{ and } \rho_1 = \sigma^2_{u1} \big/ \left(\sigma^2_{u1} + \sigma^2_{e1}\right).$$

Roberts and Roberts[84] show that where the distribution of cluster sizes differs across arms in a nested design, allowing for heteroscadasticity between arms at the therapist level by fitting model (2.3) or (2.4), or more fully with model (2.5), may give different standard errors of the intervention effect when compared to the simpler model (2.2). In their reanalysis of a trial comparing general practitioner (GP) and nurse practitioner (NP) care in primary care[101], models (2.2) and (2.5) gave noticeably different standard errors. This is explained by large numbers of GPs with smaller clusters sizes coupled with a larger ICC for GPs and a smaller standard error in the GP as opposed to the NP arm. In psychotherapy trials, cluster size is typically determined by the organisation of care and may differ systematically between arms. This may justify the use of models (2.3) to (2.5). This is in contrast with cluster randomised trials, where randomisation will ensure the average cluster size in each trial arm is similar, at least in expectation, so that models (2.2) to (2.5) give similar standard errors for the intervention effect.

If a trial has a partially nested design, each patient in the non-therapist arm could be assumed to be a cluster of size one in model (2.2). Alternatively, one might constrain the random intercept to the therapist arm using a random coefficient model[83]

$$y_l = \alpha_l + \theta t_l + v^{(2)}_{therapist(l)} t_l + e^{(1)}_l \quad (2.7)$$

While model (2.2) forces the total variance to be equal across arms, model (2.7) forces the total variance in the therapist arm to be greater than that in the control arm if σ_v^2 is positive. Roberts & Roberts[84] showed that misspecification of the patient level variance within model (2.7) can bias estimation of the between-therapist variance and hence the standard error of the intervention effect and its test size. They suggest fitting the following two-level heteroscedastic model allowing the patient level variance to be estimated separately in each arm

$$y_l = \alpha_l + \theta t_l + v_{therapist(l)}^{(2)} t_l + e_l^{(1)} + \xi_l^{(1)} t_l \quad (2.8)$$

This may be better re-parameterised as

$$y_l = \alpha_l + \theta t_l + u_{1therapist(l)}^{(2)} t_l + e_{0l}^{(1)} (1 - t_l) + e_{1l}^{(1)} t_l \quad (2.9)$$

Models (2.2) to (2.4) can also be applied for a crossed design. If model (2.2) is used (e.g. McKay *et al*[65]), the effect of the intervention is assumed to be constant across therapists and the random intercept $u_{therapist(l)}^{(2)}$ can be thought of as a stratifying effect. As such, in the special case where model (2.2) is appropriate, the crossed design may be considered a *stratified design.*

In other applications of the crossed design, between-therapist variation in the effect of the intervention is plausible because therapists may exhibit *differential* skill, expertise, competence or fidelity to the therapy model across arms. This suggests use of model (2.3)[79, 83] where σ_u^2, σ_v^2 and σ_{uv} are now all estimable[83] and σ_{uv} represents between-therapist variability in the treatment effect. Between-therapist variation in outcome σ_u^2 is easier to interpret marginally where there is a single sample of therapists, so that model (2.3) may be preferred over model (2.4)[83] in this context. If $u_{therapist(l)}^{(2)}$ is treated as a set of fixed parameters, model (2.3) is equivalent to the random-effects meta-analysis model described by Whitehead[102] (p.131) for individual-patient-data, where therapists equate to studies.

Although less intuitive, an alternative analysis for the crossed design might be a three-level model, with patients nested within therapists, and interventions both nested within and crossed with therapists

$$y_l = \alpha_l + \theta t_l + p_{therapist(l)}^{(3)} + q_{treat(l)}^{(2)} + e_l^{(1)} \quad (2.10)$$

The random effects $p^{(3)}_{therapist(l)}$ and $q^{(2)}_{treat(l)}$ have distributions $N\left(0, \sigma^2_p\right)$ and $N\left(0, \sigma^2_q\right)$ respectively and $treat(l)$ refers to the intervention within therapist received by the l^{th} patient. Here the crossed design is viewed as a specific type of nested design (see Dunn and Clark[103] (p.178) for a discussion of the relationship between crossed and nested factors). Model (2.10) is equivalent to that suggested for analyses of matched-pairs cluster randomised designs[104] where therapists correspond to strata and to that proposed for cluster-randomised crossover trials[94] with the therapists equated to clusters. In contrast to models (2.3) and (2.4), model (2.10) assumes that between-therapist variation σ^2_u is equal to $\sigma^2_p + \sigma^2_q$ for both interventions. Where σ^2_v equals $2\sigma^2_q$, constraining the covariance σ_{uv} equal to $-\sigma^2_v/2$ in model (2.3) makes the latter equivalent to model (2.10). Conversely, constraining σ^2_{0u} equal to σ^2_{1u} in model (2.4) makes the covariance σ_{01u} equal to σ^2_p in model (2.10). A final analysis option would be to treat $p^{(3)}_{therapist(l)}$ as a set of fixed effects, which is another model suggested for individual-patient-data meta-analysis[102].

The variances of $u^{(2)}_{therapist(l)}$ in model (2.3) and $p^{(3)}_{therapist(l)}$ in model (2.10) do not directly contribute to the variance of the treatment effect θ in the crossed design, but the variances of $v^{(2)}_{therapist(l)}$ and $q^{(2)}_{treat(l)}$ do. Whilst model (2.3) follows more naturally from a crossed design, model (2.10) is convenient for sample size estimation, as it partitions therapist variation into the between-therapist or marginal variance σ^2_p and the within-therapist but between-interventions variance σ^2_q.

Models (2.2) to (2.10) assume that the order in which a therapist treats their patients is unimportant. This may not be the case where therapist learning or fatigue changes a therapist's performance over time so that patient outcomes are no longer exchangeable within therapists. There may be carryover from one patient to the next, and in a crossed design from one intervention to the next, depending on the sequence of allocations. If change over time within therapists is plausible, the model adopted should ideally reflect this, perhaps through use of a more complex covariance structure. Cook *et al*[105] provide an example of a model including learning curve effects.

Models (2.2) to (2.10) can be fitted in standard multilevel software using maximum likelihood (ML), restricted maximum likelihood (REML), or alternatively within a Bayesian framework using MCMC[106]. However, small numbers of clusters (i.e. therapists) and heterogeneity of cluster sizes within arms may lead to convergence problems. It may also be difficult to fit a two-level heteroscedastic model in a Bayesian framework.

2.3.2 Power and Sample Size

Sample size estimation methods for psychotherapy trials are similar to those for cluster randomised trials. Methods for completely-randomised cluster randomised trials are directly applicable to the nested design where model (2.2) is appropriate. Donner and Klar[107] (p.57) give an asymptotic formula derived from a *z*-test for the number of clusters in each arm that can be applied in this context to choose the number of therapists $2k$

$$2k = \frac{2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\theta^2} \times \frac{\sigma^2\left(1 + (m-1)\rho_u\right)}{m} \quad (2.11)$$

where m is the number of patients treated by each therapist, and ρ_u is the ICC from model (2.2).

Where there is differential clustering (i.e. models (2.3) to (2.9)) the variance at the summary level will differ and the analysis corresponds to that of an unequal variance *t*-test[108] rather than the standard *t*-test, a point noted by Hoover[109]. This is the case particularly for partially nested designs[84]. Hoover[109] suggests using an approximation developed by DiSantostefano and Muller[110] to calculate sample size. This underestimates power where the variance is greater in the larger arm[110], which is the likely situation at a summary level in a partially nested design[84]. This also seems unnecessary given the exact method for the unequal variance *t*-test is implemented in sample size software such as nQuery Advisor[111] based on the methods described by Moser *et al*[112]. The number of therapists can therefore be estimated by taking the summary level variance to be

$$\sigma_i^2(sum) = \sigma_i^2\left(1 + (m_i - 1)\rho_i\right)/\,m_i \quad (2.12)$$

in the i^{th} intervention arm. A Stata routine[113] is available that uses the Moser *et al*[112] method in this context.

If the sample size formulae for completely-randomised cluster randomised trials are expressed in terms of normal deviates, as in (2.11), they will underestimate sample size and overestimate power where the number of clusters is small. Specifically, where the intraclass correlation coefficient is taken to be zero, this will give the same sample size and power as a trial with no clustering. In contrast, estimates based on the Moser *et al*[112] method take account of uncertainty in the cluster level variance estimates via the degrees of freedom related to the number of therapists. Using the latter method, there is a penalty even where one believes the ICC to be negligible or zero. So it is recommended that the number of therapists and patients be based on this method, rather than (2.11), to encourage designs with larger numbers of therapists, and thus greater protection of power when larger than expected values of the ICC are found.

Table 2.1 illustrates the Moser *et al*[112] method for a standardised effect size of 0.5, under model (2.2). If therapist variation is completely ignored, the total trial sample size is 128 for 80% power. If intraclass correlation is assumed to be zero, and there are only 5 therapists in each arm, power is reduced to 68%. Power would still be 80% if (2.11) had instead been used. The reduction in power is substantial with an intraclass correlation coefficient of 0.05. If the number of therapists is kept constant the increase in the required patient sample size to achieve 80% power is sizeable, with diminishing returns once the number of patients exceeds 1/ρ, a point made by Donner and Klar[92] in the context of cluster randomised trials.

Table 2.1 Sample Size and Power for a Nested Design using Moser *et al*[112] Methods

$\rho_u = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2\right)$	***Therapists in each intervention arm***	***Patients per therapist***	***Total trial patient sample size***	***Power***
		No clustering	128	80%
0	5	13	130	68%
0.025	5	13	130	56%
0.05	5	13	130	48%
		Increasing numbers of patients per therapist		
0	5	18	180	81%
0.025	5	30	300	80%
0.05	5	130	1300	80%
		Increasing numbers of therapists		
0	7	13	182	86%
0.025	8	13	208	83%
0.05	9	13	234	80%

Note: α=0.05 (two-sided); standardised effect size is 0.5

Equal allocation is the norm for randomised controlled trials, as this maximizes power for a given sample size. However, this is based on the variance of treatment means being equal for both interventions, which may not be the case in a psychotherapy trial. Where there is heteroscedasticity in the clustering effect, the power for a given sample size can be maximised by changing the allocation ratio between the intervention arms. Assuming the total variance in each arm is given by σ_{t0}^2 and σ_{t1}^2, the standard error of the intervention effect is

$$\sqrt{\frac{\sigma_{t0}^2(1+(m_0-1)\rho_0)(1+R)}{NR}+\frac{\sigma_{t1}^2(1+(m_1-1)\rho_1)(1+R)}{N}} \quad (2.13)$$

where patients are allocated in the ratio *R*:1 for intervention 0 and 1, ρ_0 and ρ_1 are the treatment-specific intraclass correlation coefficients, and *N* is the total sample size. This corrects typographical errors in Roberts[84, 113]. Assuming an asymptotic approximation, the power is maximized for a given sample size with an allocation ratio

$$R=\sqrt{\frac{\sigma_{t1}^2(1+(m_1-1)\rho_1)}{\sigma_{t0}^2(1+(m_0-1)\rho_0)}} \quad (2.14)$$

For a partially nested design this ratio simplifies to

$$\sqrt{\frac{\sigma_{t1}^2}{\sigma_{t0}^2}(1+(m_1-1)\rho_1)} \quad (2.15)$$

which is the formula given by Moerbeek and Wong[114].

Estimates of sample size and power for a crossed design under model (2.10) basically are the same as those for a matched-pairs cluster-randomised trial analysed using a matched-pairs *t*-test[107], where each therapist corresponds to a stratum. Assuming the number of patients treated by each therapist is *m* in each of two interventions (i.e. the total caseload is $2m$), Donner and Klar's[107] formula can be modified to take account of uncertainty in the variance estimates to give the number of therapists, $2k$, by

$$2k=\frac{2\left(t_{k-1,\alpha/2}+t_{k-1,\beta}^n(\varphi)\right)}{\theta^2}\times\left(\frac{\sigma_e^2}{m}+\sigma_q^2\right) \quad (2.16)$$

where t and t^n are the central and non-central t-statistics, σ_q^2 is the within-therapist variance in the intervention effect, and φ is the non-centrality parameter equal to $\theta \Big/ \sqrt{2\left(\frac{\sigma_e^2}{m}+\sigma_q^2\right)}$. The number of therapists k, being on both sides of equation (2.16), has to be found iteratively.

Implications of within-therapist variation in the intervention effect are illustrated in Table 2.2 for a standardised effect size of 0.5. Again there is a penalty for trials with small numbers of therapists, even where one believes ρ_q to be negligible. For a nested design (see Table 2.1) 208 patients and 16 (2x8) therapists would be needed where the intraclass correlation ρ_u is 0.025 under model (2.2). This reduces to 160 patients and 16 therapists where ρ_q is equal to 0.025 under model (2.10). Given the stratifying effect of ρ_p is ignored so ρ_q is likely to be smaller than ρ_u and under equal allocation model (2.2) is likely to be more powerful than models (2.3) to (2.9), the gain may be even greater. If model (2.3) is preferred over model (2.10) for the crossed design, σ_q^2 can simply be replaced by $\sigma_v^2/2$ in equation (2.16).

Table 2.2 Sample Size to Achieve 80% Power in a Crossed Design with Model (2.10)

$\rho_q = \left(\sigma_q^2\right)/\left(\sigma_q^2+\sigma_e^2\right)$	**Number of Therapists**	**Minimum number of patients per therapist to achieve 80% power**	***Total trial patient sample size***	**Power**
No therapist effect		-	128	80%
0	8	22	176	81%
0.025	8	30	240	81%
0.05	8	44	352	80%
0	12	14	168	84%
0.025	12	16	192	83%
0.05	12	18	216	81%
0	16	10	160	84%
0.025	16	10	160	80%
0.05	16	12	192	83%

Note: α=0.05 (two-sided); standardised effect size is 0.5

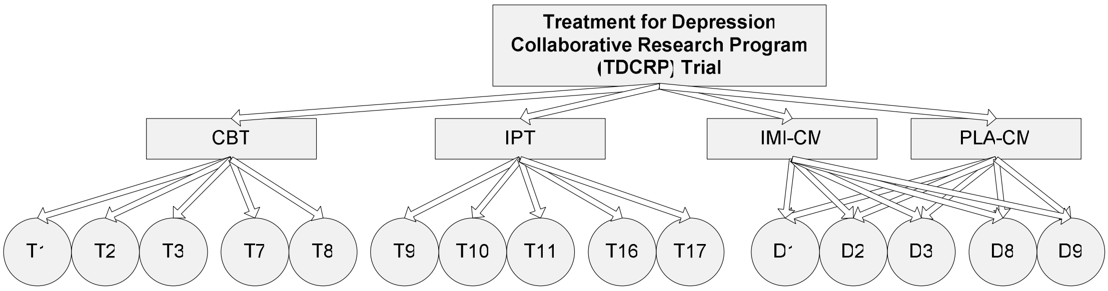
Current experience of reanalysing psychotherapy trials suggests that cluster sizes may be highly variable, with small numbers of therapists treating large numbers of patients and the remainder treating just a handful. The sample size formulae given above and illustrated in Tables 2.1 and 2.2 assume equal cluster size within arms. Power is lost as variability of cluster sizes in each arm increases, but the effect is not great provided the variance in cluster sizes within arms is relatively small[115, 116]. Unfortunately in this setting, as dedicated research therapists may often be employed to provide the bulk but not all treatment, the variation in cluster size within arms may be large. Grossly unequal cluster sizes within the arms should be avoided to prevent substantial loss of power. Experience suggests that they may also cause boundary value problems when model fitting.

2.4 More Complex Trial Designs

2.4.1 Three or More Arms

Where psychotherapy trials compare more than two interventions, the three basic trial designs (see Figure 2.2) may be found in combination. Take for example the NIMH Treatment for Depression Collaborative Research Program (TDCRP) trial[48, 117-119] that is depicted in Figure 2.3. This well-known trial individually-randomised 250 patients with a diagnosis of Depression to one of four interventions: Cognitive-Behavioural Therapy (CBT); Interpersonal Psychotherapy (IPT); Imipramine Hydrochloride plus Clinical Management (IMI-CM); or Placebo plus Clinical Management (PLA-CM). Different therapists provided CBT and IPT, such that for these interventions a nested design was used. The same therapists (psychiatrists in this case) provided the clinical management within IMI-CM and PLA-CM. As such, a crossed design was used for these interventions. Allowing for heteroscedasticity at the therapist and patient levels, model (2.5) could be extended to analyse this trial by additionally constraining the covariance σ_{uv} to the IMI-CM and PLA-CM arms.

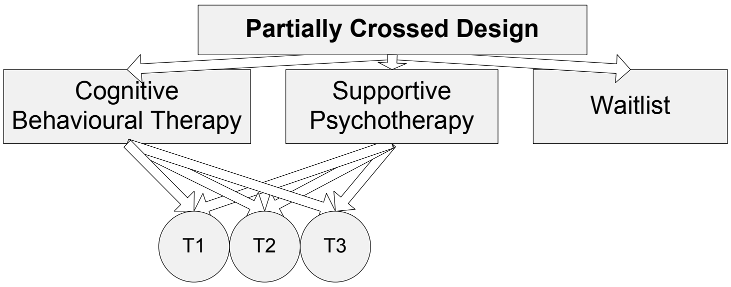
Figure 2.3 Combined Nested & Crossed Example: TDCRP trial[119]


Treatment for Depression Collaborative Research Program (TDCRP) Trial
CBT
IPT
IMI-CM
PLA-CM
T1
T2
T3
T7
T8
T9
T10
T11
T16
T17
D1
D2
D3
D8
D9

Note: T1 to T17 are therapists; D1 to D9 are doctors.

Partial clustering could also arise in the context of a crossed design where there is no therapist involvement in one (or more) arms. For example, Blanchard *et al*[120] randomised 98 patients with Posttraumatic Stress Disorder to one of three therapists and then to one of three interventions: Cognitive Behavioural Therapy (CBT); Supportive Psychotherapy (SUPPORT) or Waitlist (see Figure 2.4) using a factorial design. The same three therapists provided CBT and SUPPORT but there was no therapist involvement in the waitlist condition.

Figure 2.4 Partially Crossed Example: Blanchard *et al*[120]



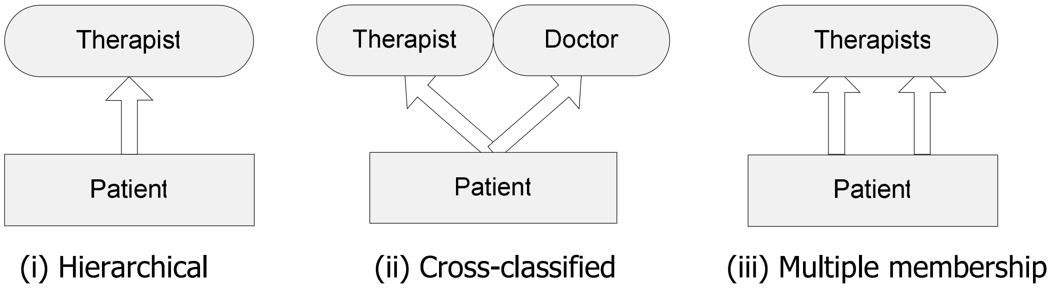
Note: T1 to T3 are therapists

Model (2.5) could be extended here to constrain both the between-therapist variance σ_u^2 and the covariance σ_{uv} to the crossed arms. However, in practice it is unlikely that this model would fit due to the small number of therapists and patients in the trial. As such, $u_{therapist(l)}^{(2)}$ might be treated as a set of fixed parameters in a meta-analysis type model.

2.4.2 Multiple Therapists-per-Patient

Up to now it has been assumed that patients receive psychotherapy from just one therapist in a simple nested design. Alternatively patients may receive psychotherapy from more than one therapist, which might occur in several ways. An intervention may be a combination of different therapeutic approaches with each approach delivered by a different type of health professional. Goldstein[95] describes this as creating a cross-classified relationship between patients and health professionals. Alternatively a course of psychotherapy might be made up of several sessions, which could be delivered by different therapists of the same type. This is an example of a multiple membership relationship[95], where sessions are not included as a specific level within the analysis. The intervention might also involve a uni- or multidisciplinary team of therapists within each session, a situation common in group-based interventions. Goldstein[95] suggested diagrams to represent simple hierarchical, cross-classified and multiple membership relationships, which can be used to illustrate the relationship between patients and therapists in a psychotherapy trial (see Figure 2.5).

Figure 2.5 Single versus Multiple Therapist-Per-Patient Designs

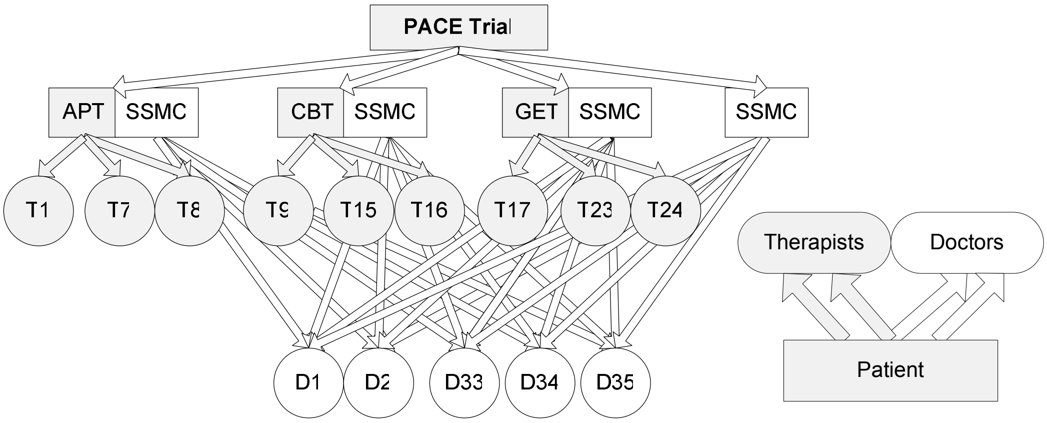


A simple hierarchical relationship is represented by a single arrow between the patient and therapist. Where a cross-classified relationship is present, a single arrow between the patient and psychotherapist *and* between the patient and doctor reflects a one-to-many relationship between patients and health professionals. This is analogous to pupils being nested both in schools and localities in educational research[95]. Where a multiple membership relationship is present, a double arrow between the patient and therapists instead reflects a one-to-many relationship between patients and therapists of the same type. In theory cross-classified and multiple membership models[121, 122] can be fitted in software such as MLwiN[123] using maximum likelihood or restricted maximum likelihood. Even if the trial generates data of sufficient quality to set up these analyses, model fitting is likely to be problematic unless sample sizes are much larger than is typically seen in psychotherapy trials. Additional levels are used to construct a hierarchy, and then constraints are added. This may be simplified in more complex situations by adopting a Bayesian framework using MCMC[106, 121].

The PACE trial[3] provides an example both of cross-classified and multiple membership relationships between patients and health professionals (see Figure 2.6). Six hundred patients with chronic fatigue syndrome/ME are being individually randomised to one of four interventions: Adaptive Pacing Therapy plus Standardised Specialist Medical Care (APT); Cognitive Behavioural Therapy plus Standardised Specialist Medical Care (CBT); Graded Exercise Therapy plus Standardised Specialist Medical Care (GET); or simply Standardised Specialist Medical Care (SSMC). The first three interventions are multi-component in nature with each component being provided by a different set of health professionals. This creates cross-classified relationships for APT/SSMC, CBT/SSMC and

GET/SSMC and a hierarchical relationship for SSMC. The fourth intervention includes only one of these components, so there is also partial nesting. For a relatively small proportion of patients, more than one therapist is involved in their care, as might be expected in clinical practice, due to staff turnover. This creates multiple membership relationships between patients and health professionals.

Figure 2.6 Multiple Therapist Per Patient Example: The PACE Trial[3]



Note: T1 to T24 are therapists; D1 to D35 are doctors. Therapists are nested; doctors are crossed; there is no therapist involvement in SSMC.

Where interventions are administered in groups, an additional source of clustering is present. The relationship between patients and groups can also be described in terms of Goldstein's[95] classification. If group membership remains fixed across the course of a trial, and patients belong only to one group, the structure is simply hierarchical with patients nested within groups, in turn nested within therapists. Many group-based intervention trials include insufficient numbers of therapists to enable all three levels to be fully taken into consideration. In these circumstances it may be more important or feasible to take account of group variation than it is to take account of therapist variation. Each group may be administered by more than one therapist, either of the same type, creating a multiple membership structure, or of a different type, creating a cross-classified structure. If the patient is involved in more than one type of group intervention as part of their treatment package (e.g. initial and maintenance groups) and membership of these groups is defined separately, the design is cross-classified. Alternatively if group membership is fluid, with patients joining or leaving each group across the course of a trial, as may be the case in rolling groups, such as Alcoholics Anonymous, one could consider this to be a multiple membership design with each group session equating to a group.

2.5 Implications for Internal Validity

When considering the degree to which the effects of interventions result from a causal association between the intervention and patient outcome, it has been suggested that effort should focus on the avoidance of four potential biases[124]. Selection biases are systematic differences in baseline characteristics between arms. Performance, attrition and detection biases arise if there is a differential effect of additional/co-interventions, loss to follow-up or outcome assessment between arms, respectively. While all four biases may be present at both the patient and therapist levels, due to the multilevel nature of the design, the focus here will be on two potential selection biases. The first relates to the method of allocating *interventions to therapists* and affects the causal interpretation of intervention effects. The second relates to the method of allocating *therapists to patients* and affects the causal interpretation of therapist variation. It is important to consider concealing allocations in both cases.

The implications of non-random, or purposive, allocation of *interventions to therapists* will depend to some extent on the research question (Figure 2.1). Any confounding of therapeutic approaches and therapist characteristics is problematic where interest is isolated to particular therapeutic approaches, or to particular therapist characteristics. Confounding of therapeutic approaches and therapist characteristics may cause little or no concern, however, where the intervention is intentionally a package, or indeed where a factorial design is used to investigate the additive *and interacting* effects of particular therapeutic approaches and particular therapist characteristics.

The focus in the PACE trial (Figure 2.6) has been on packages where each intervention component is provided by a different professional group. Broadly, APT is provided by occupational therapists, CBT by psychotherapists, GET by physiotherapists and medical care (SSMC) by doctors. Interpretation of the results will therefore be restricted to the therapeutic approaches as provided by particular professional groups within the trial. If the aim had been to make inferences regarding the therapeutic approaches in isolation the random allocation of therapeutic approaches to health professionals would have ensured average baseline comparability of therapist characteristics across therapeutic approaches.

It is not uncommon for psychotherapy researchers to note that the assignment of therapeutic approaches to therapists *ideally* should be random[7, 79, 125, 126]. Yet this is still very much an exception in practice (but see UKATT Research Team[127] and Schnurr *et al*[128, 129] for examples). Staines[126] (p.169) echoes the prevalent view, stating that "random assignment of therapists to conditions is likely to be even more difficult organizationally than random assignment of clients". It is therefore of note that Schnurr *et al*[129] (p.633) report that "there were no problems with therapists wanting to switch their assignments, or in adhering to their assigned therapy if it was not their preferred therapy". Considering the professionals perspective, Wampold and Serlin[96] suggest that non-random allocation mirrors clinical practice where therapists' have the freedom to provide their preferred therapeutic approach. However, in this case the research question relates to packages and interpretation of the results should reflect this.

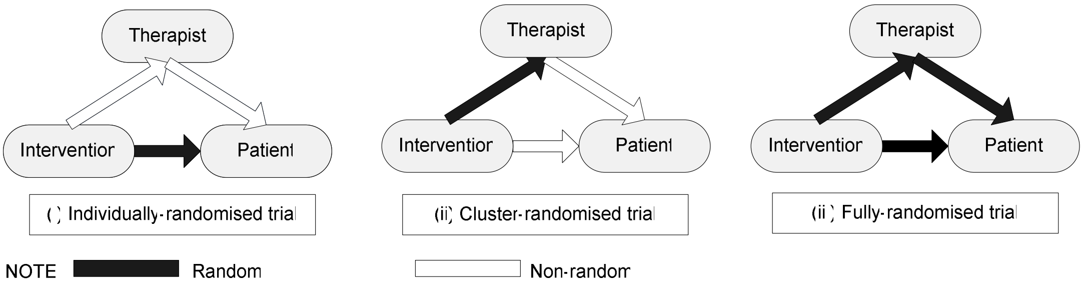
Where therapist variation is of interest in its own right, random allocation of *therapists to patients* is important if confounding of therapist variation by patient characteristics is to be avoided. Lambert[130] (p.482) suggested that "since most research does not consider each therapist to be an independent variable this procedure is rarely used", but see Brooker & Wiggins[57], Blowers *et al*[131], Borkovec *et al*[132], Durham & Turvey[133], Butler *et al*[134], Barlow *et al*[135] and Durham *et al*[136] for examples. Lambert[130] (p.482) goes on to say that random allocation "is an added burden to the research and clinical staff, but one that could be well worth the effort" as it facilitates causal interpretation of individual differences in therapist outcome. Practical difficulties relating to therapist availability at the point of randomisation may be avoided by building sufficient capacity into the design[129], something that is also needed to address the implications of small numbers of therapists for precision.

Removing confounding of therapist variation by patient characteristics could also affect the standard error of intervention effect estimates within a nested design, either by increasing *or* decreasing the size of the clustering effect. Lutz[66] suggests that where therapists specialise in specific patient groups the distribution of patient characteristics across therapists may be influenced by random assignment of therapists to patients, such that therapist variation is decreased. Conversely, where therapists are assigned suitable patients according to their level of expertise in clinical practice, Lutz[66] thought that random assignment may increase therapist variation, as differences in therapist performance may be greater when the complete spectrum of patients are included.

In some circumstances it may be either desirable or practical to maintain pre-existing therapist-patient allocations. For example, patients will often already have their own GP. In other circumstances, the geographical location will place restrictions on which therapists can be assigned to patients. For example, by embedding counsellors within general practices, counsellors are effectively allocated to general practices rather than to patients. As such, allocation of counsellors to patients is pre-determined based on the general practice patients attend. There may also be only one therapist available within a clinical service, such that therapist variation is aliased with centre effects.

A number of randomisation methods are possible both for nested and crossed designs. Three possibilities for nested designs are given in Figure 2.7. Individual randomisation refers to random allocation of treatments to patients. In Figure 2.7, the therapeutic approach, therapist characteristic or package constitute the treatment. Allocation of treatments to therapists and of therapists to patients remains non-random, such that these aspects of the design are observational. Treatment effects are hence vulnerable to confounding by therapist characteristics and their interaction with characteristics of patients. The former may not be an issue if the treatment is a therapist characteristic or package, however. Therapist variation is also susceptible to confounding by patient characteristics. This would be problematic if causal interpretation of therapist variation was desired. An alternative option might be to regard the individual therapist as the treatment, and to individually randomise treatments by allocating *individual therapists* to patients, thereby avoiding confounding both the therapeutic approach *and* therapist variation by patient characteristics. This could be attractive where the intervention is a therapist characteristic or package.

Figure 2.7 Some Possible Allocation Schemes for Nested Designs



Then again, a cluster randomised design could be employed, in which the therapeutic approach is randomly assigned to therapists. The allocation of interventions to patients might then be determined by the pre-existing assignments of therapists to patients.

Therapist variation will be vulnerable to confounding by patient characteristics in this design because the allocation of therapists to patients is observational. If unconcealed random allocation of interventions to therapists precedes the non-random allocation of therapists to patients, a recruitment bias is possible, similar to that suggested more generally for cluster randomised trials[137], due to confounding of intervention effects by patient characteristics. As such, while the clustering effect may be unaffected by the use of cluster randomisation in this context, internal validity considerations may rule against use of this design where alternatives are feasible.

Another possibility would be firstly to randomly allocate interventions to therapists and then therapists to patients using a multi-tiered experimental design[138]. The individual therapists are the experimental units in the first randomisation, and treatments in the second. If the order in which these two randomisations is performed is unimportant, the design is referred to as *composed* by Brien and Bailey[138], and has no aspect that is observational. This precludes use of dynamic allocation methods such as minimisation, where each allocation is conditional on the unit characteristics of previous allocations, which are unknown at the outset, however. Where interventions can only be randomly allocated to therapists prior to randomly allocating therapists to patients, the design is referred to as *randomised inclusive*[138]. Thus, while confounding of intervention effects by patient characteristics may be avoided if the second randomisation is concealed the effect of intermediate events (therapist training) on baseline therapist characteristics may also be an important consideration. Both multi-tiered designs may be theoretically attractive options, if therapeutic approaches are to be compared, but likewise may be difficult to implement.

Where therapeutic approaches are of specific interest it will be important to investigate baseline comparability of therapist characteristics across intervention arms[78, 139], and of patient characteristics between therapists[78], particularly if allocation of psychotherapies to therapists is non-random. Baseline comparability of patient characteristics across therapists would also be important where therapist characteristics or packages are of interest. These comparisons could be seen a check on the adequacy of randomisation, where this is appropriate, or as part of a more general assessment of the impact of confounding if aspects of the design are observational. As such, it may be important to draw upon the methods of analysis commonly used within epidemiology. Imbalance in observed therapist characteristics between arms could, in theory, be dealt with using covariate adjustment, stratification, or weighting. However, this may be difficult in practice due to limited degrees of freedom resulting from small numbers of therapists.

It may also be impossible to distinguish therapist variation from the variation in patient adherence or effect of co-interventions between-therapists and thus from performance bias when unblinded therapists provide both the intervention and any co-interventions. Similarly, therapist variation is indistinguishable from variation between interviewers in outcome assessment, and therefore detection bias, if unblinded therapists are involved in treatment *and* outcome assessment. One method of limiting the role of detection bias in therapist variation is to separate treatment from outcome assessment using blinded independent raters. The role of performance bias in therapist variation might be reduced by choosing a nested over a crossed design.

Wilkins[140] (p.4) differentiated between two types of therapist characteristic. Therapist traits, such as therapist age, gender and profession, were defined as "relatively stable and enduring across therapy and extratherapy situations". In contrast, therapist states are "situation-dependent" including therapist expertise, skill or commitment. Wilkins[140] argued that therapist traits can be controlled effectively using a crossed design, but that therapist states cannot. As such, crossed designs might be considered susceptible to performance bias. Elkin *et al*[48] adopted a nested design for CBT and IPT within the TDCRP trial (Figure 2.3) because these treatments were considered too different for therapists to provide with equal expertise, skill and commitment. It seems likely that such a design was adopted, at least in part, in an attempt to limit the potential for performance bias in a trial where blinding was not feasible. This trial can be viewed as an example of an *expertise-based trial*[141], in that much effort was put into ensuring adequate, consistent therapist expertise across interventions using a nested design.

2.6 Implications for External Validity

Martindale[79] was first to outline the statistical basis for generalisation in this context, although it has also been discussed by others[47, 96, 97, 125, 142]. Martindale suggested that random selection of patients and therapists is necessary for intervention effects to be generalised to their respective populations. He went on to argue that therapists must be a random effect in analyses for generalisations to be made on a statistical basis. At the other extreme, Siemer and Joorman[125, 142] have argued in favour of a fixed-effects approach, suggesting that inferences based on random allocation alone are possible but are limited to the local population. Crits-Christoph *et al*[97] adopted a compromise position, arguing that inferences can be made to a hypothetical population of similar therapists in the absence of random allocation. It may be concluded that the further away one moves from random selection and allocation, the less robust the basis for generalisation.

A factor is considered random if the levels included are drawn from a larger population of possible levels, and there is interest in this larger population. In contrast, a factor is considered fixed if all of the levels of interest are specifically included in the model. Arguing against a fixed effects approach, Wampold and Serlin[96] suggest that particular therapists are rarely of interest. For this reason Martindale[79] and Crits-Christoph *et al*[97] also state that there is little, if no, scientific value in treating therapists as fixed. However, it is not always possible to include therapists as a random effect, particularly in early phase trials where the number of therapists is likely to be small. Paul and Licht[47] suggest that the scientific value of a study comes primarily from its ability to provide causal inferences rather than the nature of the generalisations possible. With this in mind, it would seem reasonable to place less emphasis on generalisation during the earlier stages of development and evaluation[125], using information gathered at this stage to inform the design of a large definitive trial[81] where this is warranted.

Another aspect of external validity relates to the selection of therapists into a trial. Elkin[78] recommended reporting formal therapist eligibility criteria, suggesting that different criteria may be used to select therapists for different therapeutic approaches where packages are being compared. As such, therapists may originate from a single population or from multiple populations, depending on the particular interventions. Elkin[78] also suggested that trialists should provide parallel information on the flow of therapists through a trial, similar to that given in a CONSORT diagram for patients. This would allow readers to assess the impact of the therapist recruitment process on generalisability of results, but also staff turnover and its implications for therapist variation.

Therapists volunteering to work in a trial are likely to differ from therapists in clinical practice. They may be the therapy developers or unusually expert in early phase trials. Later on they may be younger, more enthusiastic or more committed than their counterparts in practice. This will be an issue where external validity is considered important. While considering packages or providers, Roberts[81] argued for each sample of health professionals being equally representative of its respective population within usual care. Elkin[78] recommended reporting therapist baseline characteristics to allow readers to assess the representativeness of therapist samples included in a trial. This is relevant because provision of psychotherapy is assumed to vary inside and outside of the trial setting. Recruiting therapists directly from clinical practice using broad eligibility criteria might lead to them being more representative, but it could also increase the size of the clustering effect by increasing the range of therapist expertise. If the intervention is a package or particular therapist characteristic, an assessment of the generalisability of the treatments provided in the trial would also be important[81]. This might be done by providing therapist baseline characteristics separately for each therapist sample.

2.7 Conclusions

Randomised trials of psychotherapy are characterised not only by the complexity of their interventions but also by the complexity of their designs and associated data structures. Once it is acknowledged that therapists create an additional level with the design, psychotherapy trials become part of a wider class of "multilevel" randomised trials characterised by their complex data structures. Recognition of this has been obscured by a common perception that clustering is only an issue in trials where cluster randomisation is used. Greater consideration needs to be given from the outset to the broad principles of experimental design when considering relationships between treatments and therapists and between therapists and patients. Trialists should then justify what is appropriate and feasible to address their particular research question, appreciating the consequences of adopting a particular design and analysis strategy. Clearer and more precise reporting of research questions, trial designs and therapist variation is therefore needed, as is the prospective gathering of therapist data. Even where multiple randomisations are not feasible or appropriate, it can be argued that considering them aids understanding of potential biases associated with observational aspects of a design, mirroring the view that RCTs provide a model for epidemiological studies[143].

3 A SYSTEMATIC METHODOLOGICAL REVIEW OF COCHRANE REVIEWS OF COMPARATIVE STUDIES INVOLVING PSYCHOTHERAPY

3.1 Introduction

Statistical pooling or meta-analysis of summary-data across studies can be viewed as a two-stage process in which summary statistics are first extracted from each study and then a weighted average is calculated of them[144, 145]. The summary statistic is often an odds ratio or relative risk, if outcomes are binary, or an absolute or standardised mean difference, if they are normally distributed. Weights are typically defined by the inverse of the variance of the summary statistic[146, 147] and hence are a function of the standard errors calculated within the individual studies. Observations are frequently assumed to be statistically independent of one another, with a common variance across treatment arms within studies[148]. Consequently the presence of heteroscedasticity or clustering in the studies invalidates the use of these methods, unless standard errors are estimated from analyses that properly reflect these features of the data.

The past decade has seen growing interest in the specific methodological challenges faced in the meta-analysis of randomised trials with multilevel designs. Methods have been proposed for use in pooling studies with repeated-measures[149-152], crossover[153-156] and cluster-randomised[42, 43, 157-160] designs and, recently, individually-randomised trials with inherent clustering[39]. What is common across this literature is a consideration of the impact of within-study clustering when combining data from trials with complex data structures, particularly where this has been ignored in published study analyses. While the general clustering issue is shared by all multilevel trial designs, the ensuing data structures vary in important respects. For example an assumption is usually made in relation to cluster-randomised trials that the between- and within-cluster outcome variances are equal across arms[107]. This leads to a common clustering effect across all treatment arms, indexed by an intraclass correlation coefficient. Nested therapist and group-based intervention studies are characterised by variance heterogeneity, both at the cluster (i.e. the therapist or group) and patient levels across arms, however[84]. An important corollary is that the size of any associated clustering effect is anticipated to fluctuate between treatments in such studies. It is thus necessary to consider not only the impact of clustering but also that of relaxing homoscedasticity assumptions.

Two empirical reviews have been published of the approaches taken by reviewers to the synthesis of cluster-randomised and crossover trials so far[156, 161]. Elbourne *et al*[156] found that, of the 1000 reviews published in Issue 1, 2001 of the *Cochrane Database of Systematic Reviews*, 184 included crossover trials. Of these, only Hubbert *et al*[162] adjusted for the associated clustering effects. The most frequently adopted approach was to treat crossover trials as if they had parallel-group designs. This was done by ignoring within-study clustering and including data from both treatment periods (56; 30%) or from the first period only (95; 52%). The remaining reviews either excluded crossover trials (11; 6%) or reported their findings within the text (21; 11%). A clear approach was rarely stated in the *Methods* section, and those reported varied across Cochrane Review groups.

Laopaiboon[161] searched *MEDLINE*, *EMBASE*, *Health Star*, *SCIsearch* and the *Cochrane Library* for reviews published up to 2000 that included cluster-randomised trials. Of the 25 reviews identified, 16 (64%) were found in the *Cochrane Database of Systematic Reviews*. Fifteen (60%) included more than one cluster-randomised trial, their design and unit of randomisation differing within meta-analyses for the most part. Only one review satisfactorily adjusted for associated clustering effects[163]. Fawzi *et al*[164] inflated standard errors by 30%, but did so in spite of whether the estimation method allowed for clustering. Glasziou *et al*[165] adapted the methods described by Rao and Scott[166] but appeared to regard individually-randomised trials as having one cluster per arm[161]. The most common approach was to use the published standard errors (15; 60%). This was appropriate in two reviews[167, 168], where all of the cluster-randomised trials had been analysed taking clustering into account[161]. Of the remaining reviews, 6 (24%) reported the results in the text and 1 (4%) did not clearly state what methods had been used. In no case was a rationale given, however clear, for the approach adopted[161].

The presence of additional levels in meta-analyses has potential implications for the precision, internal and external validity of the pooled treatment effect estimates. The emphasis given to precision, largely at the expense of validity, in the methodological literature is mirrored in guidance given to reviewers by the Cochrane Collaboration[169]. This was revised in May 2005 to include additional sections on crossover and cluster-randomised trials, in part as a response to this literature. The possibility of selection, performance, attrition and detection biases at multiple levels was not mentioned, nor was generalisation of treatment effects to units at all levels. What advice was given was also specific to these designs. The aim of this chapter was to explore the range, complexity and recognition of issues arising in meta-analyses of psychotherapy trials, with particular attention to the multilevel aspects. While such issues are expected to be widely applicable, psychotherapy provides a clear focus and its definition was kept deliberately broad. The absence of space constraints, along with their structured and electronic format, makes the reviews in the *Cochrane Database of Systematic Reviews* an ideal sampling frame. Extensive citation in medical journals, policy documents and practice guidelines[170] further justifies their quality being of specific interest, and hence also their likelihood of indicating current best practice.

3.2 Methods

3.2.1 Selection and Description of Systematic Reviews

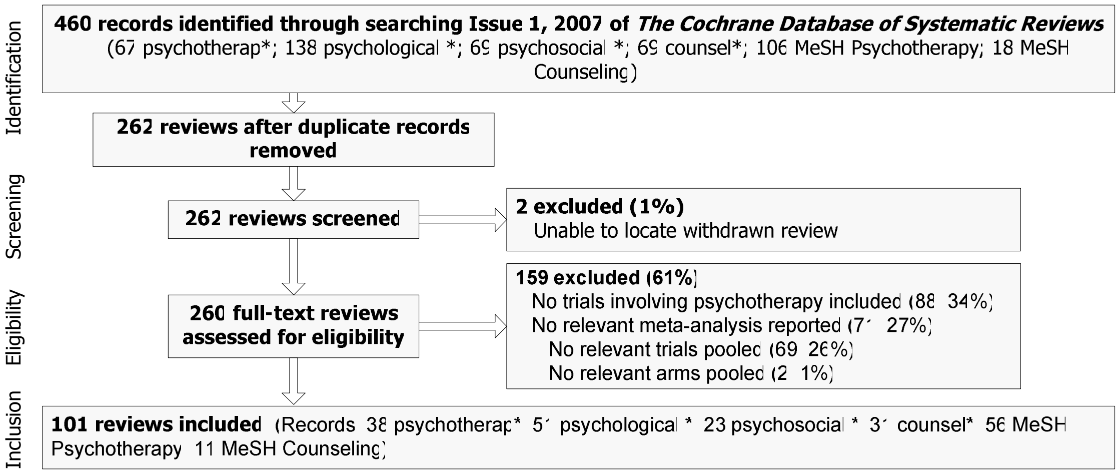
The search strategy was refined by obtaining a list of general terms used by Cochrane reviewers to refer to randomised trials involving psychotherapy. A list of specific terms was not generated because an exhaustive list would have been impractical and a selective one might have reduced the breadth of the search. One hundred and seventy-five reviews were identified in Issue 4, 2006 of the *Cochrane Database of Systematic Reviews* using keyword searches based on the terms *psychotherapy* and *psychological intervention*. A refined set of search terms was then created in light of the phrases used in the titles and abstracts of these reviews. The *Cochrane Database of Systematic Reviews* was searched again on 19th January 2007 using *psychotherap**, *psychological **, *psychosocial ** and *counsel** and the MeSH terms *Psychotherapy* and *Counseling*, including all subheadings and subtrees. All searches were restricted to reviews and were of titles, abstracts and keywords only. Once duplicate records had been removed, attempts were made to download the full-text of the remaining 262 reviews. Seven of these were classed as withdrawn with the full-text no longer available for two, leaving 260 reviews to be assessed for eligibility (see Figure 3.1).

To be eligible for inclusion, reviews had to report meta-analyses of studies involving psychotherapy. The full-text of all 260 reviews was inspected to determine whether

1. One or more study involving psychotherapy had been included
2. One or more meta-analysis involving these studies had been reported

Reasons for exclusion were noted.

Figure 3.1 Selection of Systematic Reviews


Identification
460 records identified through searching Issue 1, 2007 of *The Cochrane Database of Systematic Reviews*
(67 psychotherap*; 138 psychological *; 69 psychosocial *; 69 counsel*; 106 MeSH Psychotherapy; 18 MeSH Counseling)
262 reviews after duplicate records removed
Screening
262 reviews screened
2 excluded (1%)
Unable to locate withdrawn review
Eligibility
260 full-text reviews assessed for eligibility
159 excluded (61%)
No trials involving psychotherapy included (88 34%)
No relevant meta-analysis reported (71 27%)
No relevant trials pooled (69 26%)
No relevant arms pooled (2 1%)
Inclusion
101 reviews included (Records 38 psychotherap* 51 psychological * 23 psychosocial * 31 counsel* 56 MeSH Psychotherapy 11 MeSH Counseling)

A random 10% of the reviews were independently evaluated by a second reviewer (TC), with disagreements resolved by discussion and reference to the full-text. A further 15 borderline or grey-area reviews were also discussed in detail. The eligibility criteria were clarified accordingly:

i) Any experimental or control intervention referred to in any section of the review as including *psychotherapy* or *counselling* or a *psychological, psychotherapeutic, psychosocial* or *behavioural* component was accepted as involving psychotherapy except where it was made clear that there was no therapist involvement
ii) Interventions directed at care providers or organisations were excluded as not involving psychotherapy
iii) An intervention accepted as involving psychotherapy in one review was deemed to involve psychotherapy in all reviews
iv) Psycho-education was excluded as a form of education
v) Psychotherapy could be a primary or a co-intervention
vi) Studies could be randomised or non-randomised
vii) Relevant meta-analyses could include both relevant and non-relevant studies but must include one or more relevant treatment arms

Eligibility assessment was then repeated for all 260 reviews using the refined criteria. Decisions regarding 63 (62%) of the included and 65 (41%) of the excluded reviews could have been made using the abstract alone. In no review were studies excluded or reported qualitatively due to issues surrounding therapist variation.

Table 3.1 Characteristics of Systematic Reviews Included by their Focus

	Focus of Systematic Review				
	Psychotherapy			Other	
	General	Specific	Therapist Characteristics	Wider	Different
Search Term, n (% total)					
Psychotherap*	17 (81)	12 (36)	2 (40)	4 (19)	3 (14)
Psychological *	18 (86)	16 (48)	1 (20)	7 (33)	9 (43)
Psychosocial *	6 (29)	11 (33)	1 (20)	2 (10)	3 (14)
Counsel*	4 (19)	6 (18)	5 (100)	8 (38)	8 (38)
MeSH Psychotherapy	16 (76)	23 (70)	2 (40)	12 (57)	3 (14)
MeSH Counseling	0 (0)	3 (9)	4 (80)	3 (14)	1 (5)
Cochrane Review Group, n (% total)					
Depression Anxiety & Neurosis	6 (29)	6 (18)	2 (40)	5 (24)	2 (10)
Tobacco Addiction	-	5 (15)	2 (40)	2 (10)	4 (19)
Schizophrenia	1 (5)	8 (24)	-	1 (5)	1 (5)
Developmental Psychosocial & Learning Problems	1 (5)	6 (18)	-	-	1 (5)
Drugs & Alcohol	2 (10)	-	-	1 (5)	5 (24)
Incontinence	-	2 (6)	-	3 (14)	1 (5)
Heart	2 (10)	1 (3)	-	-	2 (10)
Pregnancy & Childbirth	1 (5)	-	-	3 (14)	1 (5)
Airways	2 (10)	-	-	-	1 (5)
Cystic Fibrosis & Genetic Disorders	1 (5)	1 (3)	-	-	-
Dementia Anxiety & Neurosis	-	2 (6)	-	-	-
HIV/AIDS	-	-	-	1 (5)	1 (5)
Metabolic & Endocrine Disorders	1 (5)	-	-	1 (5)	-
Pain Palliative & Supportive Care	2 (10)	-	-	-	-
Stroke	-	-	-	2 (10)	-
Back	-	1 (3)	-	-	-
Breast Cancer	1 (5)	-	-	-	-
Consumers & Communication	-	-	-	-	1 (5)
Ear Nose & Throat Disorders	-	1 (3)	-	-	-
Effective Practice & Organisation of Care	-	-	1 (20)	-	-
Gynaecological Cancer	-	-	-	1 (5)	-
Injuries	-	-	-	1 (5)	-
Multiple Sclerosis	1 (5)	-	-	-	-
Musculoskeletal	-	-	-	-	1 (5)
Updated since Cochrane Handbook Version 4.2.5 released, n (% total)					
Yes – Substantial	7 (33)	17 (52)	1 (20)	5 (24)	7 (33)
Yes – Not Substantial	9 (43)	4 (12)	1 (20)	7 (33)	5 (24)
Trials within Reviews, n (% total)					
All involve Psychotherapy	18 (86)	32 (97)	4 (80)	3 (14)	1 (5)
One or more involve Group-Based Interventions	20 (95)	26 (79)	5 (100)	15 (71)	17 (81)
One or more involve Cluster Randomisation	9 (43)	10 (30)	4 (80)	9 (43)	7 (33)
One or more are Crossover Trials	1 (5)	5 (15)	0 (0)	2 (10)	4 (19)
Meta-Analyses within Reviews, n (% total)					
All involve Psychotherapy	20 (95)	33 (100)	5 (100)	11 (52)	9 (43)
One or more involve Group-Based Interventions	18 (86)	24 (73)	4 (80)	12 (57)	14 (67)
One or more involve Cluster Randomised Trials	8 (38)	9 (27)	3 (60)	4 (19)	6 (29)
One or more involve Crossover Trials	1 (5)	4 (12)	0 (0)	0 (0)	4 (19)
TOTAL (Unique Systematic Reviews)	**21**	**33**	**5**	**21**	**21**

Table 3.1 summarises the characteristics of these reviews by their focus. Of the 101 reviews included, 59 were psychotherapy focused, while the remaining 42 had a focus other than psychotherapy. Of the former, 21 focused on psychotherapy generally[171-191], 33 on a specific form of psychotherapy[37, 192-223], and 5 on characteristics of the care providers[224-228]. Of the latter, 21 had a wider focus[229-249], including psychotherapy but also other interventions, and 21 had a different focus, where psychotherapy could be regarded as more of an aside[250-270]. It is evident that the search terms were differentially effective for selecting some types of review and that a range of terms were required to capture the heterogeneity present. It is also clear that therapist variation is an issue for a broad set of reviews, spanning 24 (46%) of the Cochrane Review groups. This was especially true for the Depression, Anxiety & Neurosis group and for the Tobacco Addiction and Schizophrenia groups, as these three groups accounted for almost half of the reviews included. Sixty-three had been updated since the release of guidance on handling clustering effects within cluster-randomised and crossover trials, with 37 of these making substantial amendments. As such, general awareness of clustering was anticipated amongst a sizeable proportion.

There was psychotherapy involvement in every included study within 58 reviews and in every reported meta-analysis in 78 reviews. The prevalence of studies with group-based intervention components was also high, with 83 reviews including at least one study and 72 reporting meta-analyses involving such studies. This is in contrast to the prevalence of cluster-randomised and crossover trials in these reviews. Only 39 included cluster-randomised trials, with 30 including one in a reported meta-analysis. Likewise, only 12 reviews included crossover trials and only 9 reviews reported meta-analyses involving them.

Table 3.2 Types of Outcomes in Reported Meta-Analyses

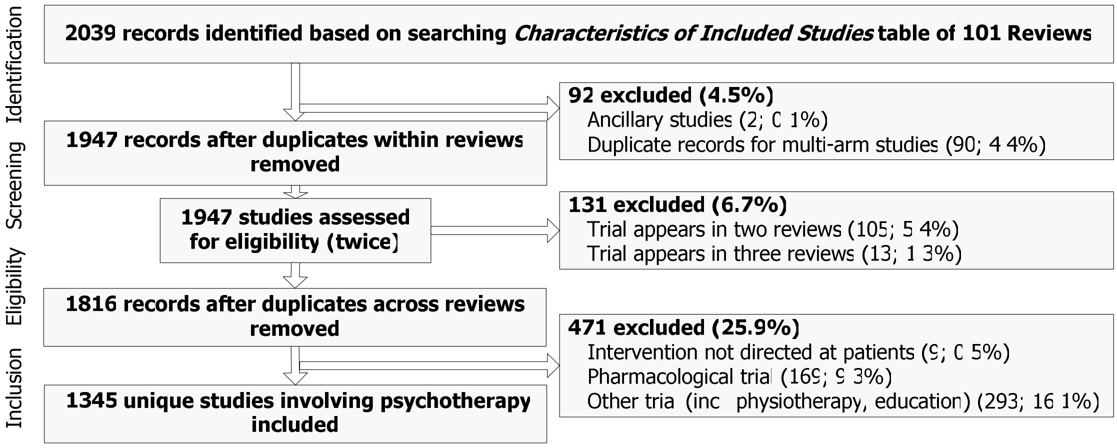
Outcomes, n (% total)	**Focus of Systematic Review**		**Overall**
	Psychotherapy	**Other**	
All Binary	**15 (25)**	**22 (52)**	**37**
All Relative Risks	6 (10)	10 (24)	16
All Odds Ratios	6 (10)	7 (17)	13
All Peto Odds Ratios	2 (3)	5 (12)	7
Mixed Binary	1 (2)	0 (0)	1
All Continuous	**18 (31)**	**7 (17)**	**25**
All Weighted Mean Differences	7 (12)	3 (7)	10
All Standardised Mean Differences	8 (14)	3 (7)	11
Mixed Continuous	3 (5)	1 (2)	4
Mixed Binary & Continuous	**26 (44)**	**13 (31)**	**39**
TOTAL	**59**	**42**	**101**

The types of outcomes reported in the meta-analyses are given by the review's focus in Table 3.2. A higher percentage of those focused on psychotherapy only reported continuous outcomes, (31% vs. 17%), while the reverse was apparent for the binary outcomes, (25% vs. 52%). A sizeable percentage reported a mixture of continuous and binary outcomes.

3.2.2 Selection and Description of Studies within Cochrane Reviews

The *Characteristics of Included Studies* table in the appendix of each included review was used to define the sample of studies. Once duplicate records within reviews had been removed, the remaining 1947 records, and associated full-text, were evaluated for evidence of psychotherapy involvement using the criteria given above. Figure 3.2 summarises this process.

Figure 3.2 Selection of Studies Involving Psychotherapy in Reviews


Identification
Screening
Eligibility
Inclusion
2039 records identified based on searching *Characteristics of Included Studies* table of 101 Reviews
92 excluded (4.5%)
Ancillary studies (2; 0.1%)
Duplicate records for multi-arm studies (90; 4.4%)
1947 records after duplicates within reviews removed
1947 studies assessed for eligibility (twice)
131 excluded (6.7%)
Trial appears in two reviews (105; 5.4%)
Trial appears in three reviews (13; 1.3%)
1816 records after duplicates across reviews removed
471 excluded (25.9%)
Intervention not directed at patients (9; 0.5%)
Pharmacological trial (169; 9.3%)
Other trial (inc. physiotherapy, education) (293; 16.1%)
1345 unique studies involving psychotherapy included

Psychotherapy involvement was assessed twice to increase consistency of ratings. It became apparent, when comparing ratings across reviews, that reporting of study arms was selective for multi-arm studies in some reviews. As a consequence, studies were regarded as involving psychotherapy if any rating was positive for a study. As the issues arising from therapist variation are equally applicable to education and physiotherapy, it was evident they were relevant to the majority of studies included in the 101 reviews.

Table 3.3 summarises some of the characteristics of the study sample. While studies described as non-randomised were included, they formed less than 5% of the total sample. There was some indication that the prevalence of multi-arm parallel-group designs is higher for studies with psychotherapy involvement, and that the reverse is true for crossover trials. The overall percentage of cluster-randomised and crossover trials in the sample was low at 6% and 2% respectively, compared to 74% involving psychotherapy. Owing to the variation in the number of arms per study, the study sample sizes are given per arm with crossover trials counted as having a single arm. Sample sizes were not reported consistently so the number randomised or enrolled is given, if possible, otherwise the number reflects those analysed. It can be seen that over half of the studies had sample sizes of 50 or less per arm. There is also some suggestion that sample sizes are smaller where there is psychotherapy involvement in a study.

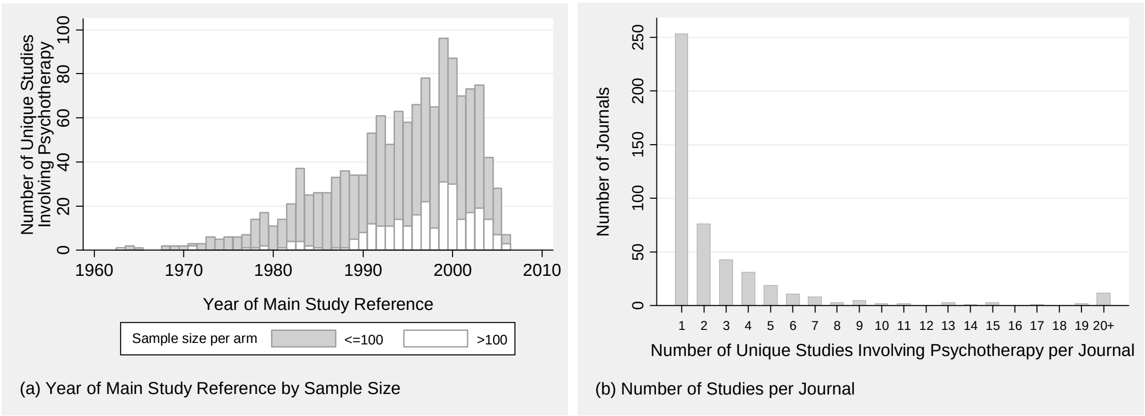
Table 3.3 Characteristics of Studies across Reviews

	Studies involving Psychotherapy	**All Studies**
Study Design, n (% total)		
Randomised	1293 (96.1)	1733 (95.4)
3+ Arm Parallel-Group	394 (29.3)	489 (26.9)
Crossover	9 (0.7)	35 (1.9)
Cluster-Randomised	81 (6.0)	111 (6.1)
Sample Size per Arm, n (% total)		
<= 10	128 (9.5)	153 (8.4)
11 to 20	260 (19.3)	340 (18.7)
21 to 50	469 (34.9)	597 (32.9)
51 to 100	212 (15.8)	293 (16.2)
101 to 200	124 (9.2)	196 (10.8)
201 to 500	91 (6.8)	138 (7.6)
501+	61 (4.5)	97 (5.3)
TOTAL (Unique Studies)	**1345**	**1816**

Note: Sample size was not reported for two cluster-randomised studies which did not involve psychotherapy

Included studies in each review were listed together with their references. Where more than one citation was given per study, reviewers often starred one to indicate it as the main citation. For studies with psychotherapy involvement, the source and year of the main citation was recorded along with the source of any other citations. Figure 3.3(a) gives the age distribution of the included studies by the size of their sample per arm. The number of studies increases over time with a time lag indicated in the inclusion of more recent studies. The number of larger studies also appears to increase, although the proportion of such trials does not do so noticeably. The 1345 unique studies involving psychotherapy were published in 475 different journals. The distribution of the number of studies per journal is given in Figure 3.3(b). It can be seen that the majority of journals were recorded as the source of only one study in the sample. Only 12 journals were the source of 20 or more studies (see Table 3.4).

Figure 3.3 Publication of Studies Involving Psychotherapy



(a) Year of Main Study Reference by Sample Size

(b) Number of Studies per Journal

The *Journal of Consulting and Clinical Psychology* was the most frequently cited journal, being the source of more studies than four major general medical journals combined. It was responsible for less than 10% of the study sample, however. Four major general psychiatry journals were also only the source of a tenth of the study sample, suggesting that single journals and groups of high impact journals do not serve as good sampling frames for evaluations of studies involving psychotherapy.

Table 3.4 Journals Publishing Studies Involving Psychotherapy

	N studies (%) **Any Reference**	**N studies (%)** **Main Reference**
12 Most Frequently Cited Journals:		
Journal of Consulting and Clinical Psychology	127 (9.4)	114 (8.5)
British Journal of Psychiatry	57 (4.2)	44 (3.3)
Archives of General Psychiatry	50 (3.7)	42 (3.1)
British Medical Journal	42 (3.1)	33 (2.5)
Behaviour Research and Therapy	39 (2.9)	34 (2.5)
American Journal of Psychiatry	39 (2.9)	25 (1.9)
Behavior Therapy	32 (2.4)	29 (2.2)
Addictive Behaviors	30 (2.2)	26 (1.9)
Preventive Medicine	29 (2.2)	20 (1.5)
Psychological Medicine	24 (1.8)	19 (1.4)
American Journal of Public Health	21 (1.6)	18 (1.3)
Archives of Internal Medicine	20 (1.5)	19 (1.4)
Four Major General Medical Journals	90 (6.7)	75 (5.6)
Four Major General Psychiatry Journals	153 (11.4)	130 (9.7)

Note: The four major general medical journals were BMJ, Lancet, JAMA and NEJM; the four major general psychiatry journals were American Journal of Psychiatry, Archives of General Psychiatry, British Journal of Psychiatry and Psychological Medicine.

3.2.3 Data Extraction and Analysis

A data extraction sheet was used to record information on reviewers' awareness of the precision, internal and external validity implications of therapist variation, and of other sources of clustering effects, for each included review. Quotes were extracted as comprehensively as possible in support of positive ratings, and supplemented by systematic keyword searches of the electronic full-text. This data was independently extracted for a random 10% of reviews by a second reviewer (CR). Disagreements were again resolved by discussion and reference to the full-text. During this process it became clear that reviewers' reporting of data structures, cluster sizes, intracluster correlations, multiple randomisations, therapist number and characteristics was more appropriately recorded at the study level. Therefore, the *Characteristics of Included Studies* table was systematically searched for this information, which was entered directly into a database designed for the purpose. This was then supplemented by information taken from the data extraction sheets and keyword searches of the text. Issues arising at this stage were discussed with a second reviewer (MD), and coding schemes refined accordingly. The extraction process was then repeated at the study level to improve consistency of ratings across reviews. Analyses of the data were descriptive, with themes applied to qualitative data where this helped to structure this information. Where the number of quotes is large, example quotes are provided in the text with the full quotes given in the appendix (see Section 3.9).

3.3 Recognition of Therapist Variability

Seventeen of the 101 reviews contained some general reference to variation in patient outcomes between therapists, with all but two of these having a psychotherapy focus. As can be seen from the examples in Box 3.1, these are parenthetical. A number made reference to conventional explanations for the consequences of therapist variability (see Box 3.2). Again, all but two of these reviews focused on psychotherapy. These were classified as relating to (i) treatment standardisation or therapist competence; (ii) placebo or non-specific effects; and (iii) process research. In each case the therapist is viewed as an aside. Finally, 11 reviews explored therapist characteristics as sources of between-arm or between-study heterogeneity (see Box 3.3 for examples), including the five reviews focusing on the former[224-228]. Here, there is specific interest in the therapists.

It is clear from Boxes 3.1-3.3, and other references to treatment standardisation in the reviews, that there is widespread acknowledgment amongst Cochrane reviewers of the presence of therapist variability. A deeper consideration of its impact on study designs and analyses, in pooled comparisons of arms defined by therapist characteristics, and in the analysis of subgroups of studies is also evident.

Box 3.1 Examples of General Awareness of the Presence of Therapist Variability

Ebrahim S, Beswick A, Burke M, Davey SG. Multiple risk factor interventions for primary prevention of coronary heart disease[260]:
Methodological Quality "It is likely that the quality of the intervention, in terms of...person carrying out activities...will determine the impact of intervention."

Hajek P, Stead LF. Aversive smoking for smoking cessation[192]:
Methodological Quality "it is generally believed that the same method can achieve different results when applied by different therapists."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Background "a number of methodological challenges were evident...The type of intervention, content, theoretical basis, intensity, duration, length of each session, whether one-to-one or in groups can vary, as can the profession and experience of the person delivering the intervention, and the location. This heterogeneity could make it difficult to combine the results from different studies."

Box 3.2 Conventional Explanations for the Consequences of Therapist Variability

TREATMENT STANDARDISATION AND THERAPIST COMPETENCY

Hackett ML, Anderson CS, House AO. Interventions for treating depression after stroke[239]:
Discussion "For psychotherapy trials, there is also good evidence that efficacy is linked to delivery of an adequate exposure to the intervention. This means that therapists should be trained and supervised in the therapy they are delivering, and use a standardised, pre-specified, framework for therapy. To achieve this in psychotherapy trials, the therapy is often manualised and the research therapists are trained and supervised in the use of the manual. Success in brief therapy is linked to adherence to the therapeutic model as well as to the therapists' characteristics. Future stroke psychotherapy trials should also adhere to these standard psychotherapy research guidelines if there is to be any probability of demonstrating consistency and response."

"NON-SPECIFIC" OR PLACEBO EFFECTS

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Discussion "Psychological interventions are complex in that they usually consist of a number of different elements. Some of these elements will be active ingredients specifically included because they are based on psychological theory. Other elements may not be specific to psychological interventions and may be common to many different types of intervention (such as interacting with other people with MS in a group). Still other elements will be specific to individual therapists (for example the therapist's experience and enthusiasm, and the way the therapist interacts with the client)."

"PROCESS" RESEARCH

Barlow J, Coren E, Stewart-Brown SSB. Parent-training programmes for improving maternal psychosocial health[216]:
Conclusion "There is very little research available to date addressing the role of 'process' factors, such as the way in which the programme is delivered, in producing positive outcomes with regard to parental functioning. However, it seems likely that the group facilitator/leader has an important part to play in helping parents not only to persist with a particular programme (Frankel 1992), but in facilitating an atmosphere of openness and trust between the participating parents, and in helping parents to feel respected, understood, and supported. Group leaders can play an important role in modelling attributes such as empathy, honesty and respect, and personal qualities such as a sense of humour, enthusiasm, flexibility, and warmth. The absence of data on process factors in the studies that were included in this review precludes the possibility of assessing to what extent the lack of positive change, where this occurred, was due to the content of the programme or its delivery...factors [that] are specific to particular training programmes or non-specific factors such as group-leader qualities...However, as discussed above, the literature lacks any discussion of the process of service delivery or its impact on psychosocial outcomes. Future research would benefit from some consideration of the impact of such factors on the outcomes recorded."

Box 3.3 Examples of Therapist Variation as Arm or Study Characteristics

REVIEWS FOCUSING ON THERAPIST CHARACTERISTICS

Rice VH, Stead LF. Nursing interventions for smoking cessation[225]:
Background "The aim of this review is to examine and summarize randomized clinical trials where nursing provided smoking cessation interventions. The review therefore focuses on the nurse as the intervention provider, rather than on a particular type of intervention." *Discussion* "The US Public Health Service clinical practice guideline 'Treating Tobacco Use and Dependence' (AHRQ 2000) used logistic regression to estimate efficacy for interventions delivered by different types of providers. Their analysis did not distinguish among the non-physician medical healthcare providers, so that dentists, health counsellors, and pharmacists were included with nurses. The guideline concluded that these providers were effective (Table 15, OR 1.7, 95% CI 1.3 to 2.1)." *Conclusion* "The evidence suggests that brief interventions from nurses who combine smoking cessation work with other duties are less effective than longer interventions with multiple contacts, delivered by nurses with a role in health promotion or cardiac rehabilitation."

EXAMPLE FROM OTHER REVIEWS

Hodnett ED, Fredericks S. Support during pregnancy for women at increased risk of low birthweight babies[249]:
Background "Debates have arisen regarding the relative benefits of 'professional' versus 'peer' support. Social support from a woman in one's community, who has a similar socioeconomic background and is experiencing similar life stresses, may be qualitatively different from support from a healthcare professional, who has broad professional knowledge and experience, but may not share the same socioeconomic background or life concerns, and who often provides other professional services as well as support. This Review includes studies of support by providers with varying backgrounds and qualifications." *Objectives* "Secondary objectives were to determine whether effectiveness of support was mediated by...type of provider (a healthcare professional or a lay woman)." *Methods* "A subgroup analysis was planned to compare support provided by lay women versus support by healthcare professionals, because another Review of support for childbearing women (Hodnett 2003) found differences in the effects of support by hospital staff (nurses, midwives) versus support by lay women." *Results* "Because there was only one trial in which the support was provided by lay women (Spencer 1989), and in another trial the support was provided by a multidisciplinary team that included lay women (McLaughlin 1992), the planned subgroup analysis was not performed. However, the results of these two trials were remarkably consistent with those of the other trials."

3.4 Recognition of the Implications for Precision

3.4.1 Reporting of Study Designs and Associated Data Structures

Study designs were principally reported in the *Description of Studies* section of the main text, the *Characteristics of Included Studies* table in the appendix, or in both of these. As details of the designs were not always available or reported in the reviews, designs were recorded along with whether they had been reported explicitly, implicitly or assumed. In contrast to other designs, crossover trials tended to be over-reported. There was evidence of this in five of the reviews[177, 217, 221, 244, 264], where a waitlist control receiving the intervention at the end of the trial and then followed up was confused with random allocation of treatment sequences. Reports were explicit in all cases. If there was no indication of whether a study had a crossover or parallel-group design the latter was assumed. Table 3.5 summarises reporting of cluster-randomised, group-based and therapist designs by the focus of the review.

Table 3.5 Overall Reporting of Study Designs

Designs, n reviews (% relevant total)	Assumed for all studies	Implicit for all studies	Explicit for all studies	Either Assumed or Implicit for studies	Either Assumed or Explicit for studies	Either Implicit or Explicit for studies	Either Assumed, Implicit or Explicit for studies	**Total**
Review Focus: Psychotherapy								
Cluster Randomisation[1]	0 (0)	9 (39)	10 (43)	1 (4)	0 (0)	2 (9)	1 (4)	**23 (100)**
Individual or Group-Based Interventions[1]	6 (10)	0 (0)	15 (25)	1 (2)	22 (37)	6 (10)	9 (15)	**59 (100)**
Therapists Nested or Crossed With Interventions[2]	28 (47)	0 (0)	0 (0)	9 (15)	8 (14)	0 (0)	14 (24)	**59 (100)**
Single or Multiple Therapists per Patient[2]	23 (39)	1 (2)	2 (3)	7 (12)	11 (19)	2 (3)	13 (22)	**59 (100)**
Review Focus: Other								
Cluster Randomisation[1]	0 (0)	5 (31)	6 (38)	1 (6)	0 (0)	3 (19)	1 (6)	**16 (100)**
Individual or Group-Based Interventions[1]	8 (19)	0 (0)	2 (5)	1 (2)	20 (48)	1 (2)	10 (24)	**42 (100)**
Therapists Nested or Crossed With Interventions[2]	24 (57)	0 (0)	0 (0)	6 (14)	3 (7)	1 (2)	8 (19)	**42 (100)**
Single or Multiple Therapists per Patient[2]	17 (40)	1 (2)	0 (0)	7 (17)	5 (12)	1 (2)	11 (26)	**42 (100)**

[1] Based on all 1816 unique studies; [2] Restricted to the 1345 unique studies involving psychotherapy

Cluster randomisation was explicitly reported for one or more studies in 59% (10+2+1 +6+3+1=23/39) of the reviews, implicitly reported for one or more studies in 59% (9 +1+2+1+5+1+3+1=23/39) and assumed, using information reported elsewhere, for one or more studies in 10% (1+1+1+1=4/39). Given the small number of reviews in which cluster-randomised trials were observed, there is no evidence to suggest that the level of reporting differed according to the review focus. Seven reviews including one or more cluster-randomised trial made implicit reference to the co-assignment of partners, friends or households[183, 187, 192, 207, 247, 249, 261] with 4 of these reviews relating to smoking cessation. Together with more overt references, where randomisation was described as being by a unit other than the patient, these reports were implicit to the extent that reviewers did not refer to designs as cluster-randomised and it was not always clear that they considered them as such. Even where the presence of cluster randomisation was explicit, the specific design, be it completely-randomised, matched, stratified, or crossed, was reported systematically in only 2 reviews[227, 232]. The unit of randomisation was reported more frequently, but not universally, with 12% (13/111) of cluster-randomised studies being reported by reviewers without this information.

In contrast to the reporting practices for crossover and cluster-randomised designs, those for therapist and group-based interventions did not appear to be reflective of reviewers' awareness of the implications of study design for precision. Instead, these aspects tended to be regarded as part of the description of the intervention. Table 3.5 indicates that group- versus individually-based formats were reported explicitly for one or more studies in 84% (15+22+6+9+2+20+1+10=85/101) of the reviews, implicitly for one or more studies in 28% (1+6+9+1+1+ 10=28/101) and assumed for one or more studies in 76% (6+1+22+9+8+1+20+10=77/101). Although the prevalence of group-based designs may be under-estimated, the information provided appeared to be sufficient to make this judgement, perhaps with the exception of co-interventions. It was not sufficient to determine relationships between interventions and therapists and between therapists and patients, however. These were explicitly reported in 34% (8+14+3+1+8=34/101) and 45%(2+11+2+13+5+1+11=45/101) of the reviews and then only for a proportion of studies involving psychotherapy. Assumptions for these design features were prone to considerable error, being based primarily on the type of control group or use of co-interventions for the former, the nature of the intervention in the latter, and reporting of therapist characteristics in both cases. Better reporting of therapist and group-based study designs was indicated for the reviews with a

Table 3.6 Reporting of Group- and Therapist-Based Interventions

Designs, n reviews (% relevant total)	Assumed for all studies	Implicit for all studies	Explicit for all studies	Assumed or Implicit for studies	Assumed or Explicit for studies	Implicit or Explicit for studies	Assumed, Implicit or Explicit for studies	**Total**
Individual versus Group-Based Interventions[1]								
Individual	41 (44)	2 (2)	16 (17)	1 (1)	31 (33)	0 (0)	3 (3)	**94 (100)**
Group	1 (1)	4 (5)	56 (72)	0 (0)	1 (1)	10 (13)	6 (8)	**78 (100)**
Individual & Group	0 (0)	4 (11)	27 (77)	0 (0)	1 (3)	3 (9)	0 (0)	**35 (100)**
Relationship between Interventions and Therapists[2]								
Nested	51 (55)	1 (1)	0 (0)	15 (16)	8 (9)	2 (2)	16 (17)	**93 (100)**
Crossed	58 (76)	0 (0)	5 (7)	3 (4)	7 (9)	0 (0)	3 (4)	**76 (100)**
Nested & Crossed	21 (72)	4 (14)	1 (3)	3 (10)	0 (0)	0 (0)	0 (0)	**29 (100)**
Relationship between Therapists and Patients[2]								
Single	49 (51)	3 (3)	1 (1)	19 (20)	10 (10)	1 (1)	14 (14)	**97 (100)**
Multiple: Simultaneous	3 (11)	2 (7)	13 (46)	1 (4)	3 (11)	3 (11)	3 (11)	**28 (100)**
Multiple: Parallel	28 (58)	1 (2)	6 (13)	4 (8)	7 (15)	2 (4)	0 (0)	**48 (100)**

[1] Based on all 1816 unique studies; [2] Restricted to the 1345 unique studies involving psychotherapy; In each case, the most complex structure is given per study and then the presence of one or more study with each structure is reported; Individual & group and nested & crossed imply multiple intervention components with different structures; Simultaneous multiple refers to multiple therapists per patients within an intervention component; Parallel multiple refers to one or more therapists for each of multiple intervention components; Where no relevant information was given regarding a study design, the default was to assume an individual intervention, nested therapists and a single therapist-per-patient for that study.

psychotherapy focus but reporting was poor in all cases.

Table 3.6 summarises the reporting of therapist and group-based data structures for the specific designs. The intervention format and the relationship between therapists and patients differed across treatment arms in many of the studies. Where this was so, the most complex data structure was recorded. While their prevalence remains largely uncertain, it is clear that all the conceivable designs were present across the reviews. A mixture of designs was evident within reviews, and hence also potentially in the meta-analyses. This complexity and variety has repercussions for assessments of how the potential clustering effects are handled in study reports and for methods that allow for such effects in meta-analyses.

Partial nesting or crossing of therapists and groups with interventions occurs when clustering is restricted to a subset of treatment arms within a study. This was not explicitly reported in any of the reviews. It was however indicated by the nature of the control arms in 93% (94/101) reviews for therapist designs and in 81% (67/83) of the reviews for group-based designs. Table 3.7 summarises the extent to which partial clustering was indicated by the review focus.

Table 3.7 Implicit Reporting of Partial Clustering within Cochrane Reviews

	Focus of Systematic Review					
	Psychotherapy			**Other**		
Partial Clustering Indicated	General	Specific	Therapist Characteristics	Wider	Different	**Overall**
Therapist Variation, n (% total)[1]						
All Relevant Studies	2 (10)	5 (15)	-	5 (24)	4 (19)	16 (16)
Some Relevant Studies	19 (90)	27 (82)	5 (100)	16 (76)	11 (52)	78 (77)
No Relevant Studies	-	1 (3)	-	-	6 (29)	7 (7)
TOTAL	**21**	**33**	**5**	**21**	**21**	**101**
Group Variation, n (% total)[2]						
All Studies	6 (30)	9 (35)	3 (60)	3 (20)	3 (18)	24 (29)
Some Studies	12 (60)	15 (58)	2 (40)	9 (60)	5 (29)	43 (52)
No Studies	2 (10)	2 (8)	-	3 (20)	9 (53)	16 (19)
TOTAL	**20**	**26**	**5**	**15**	**17**	**83**

[1] Restricted to the 1345 unique studies involving psychotherapy; [2] Based on all 1816 unique studies

It is evident that the absence of such studies was a particular feature of reviews with a different focus. This was because the therapist and group involvement in these reviews tended to be in the form of co-interventions representing standard care. Where reviews include multi-arm studies, the impact of partial clustering on meta-analyses depends on which arms are selected. Therefore, while the impact is possibly overstated, it is clear that partial clustering is a common issue for Cochrane reviews and meta-analyses of studies involving psychotherapy.

3.4.2 Reporting of the Extent and Handling of Study Clustering Effects

Reviewers reporting of the extent and handling of study-level clustering effects more overtly reflected their awareness of the precision implications. Table 3.8 summarises reporting of cluster sizes and intraclass correlation coefficients (ICCs) by the source of the potential clustering effect.

Table 3.8 Reporting of Cluster Sizes and Intraclass Correlations

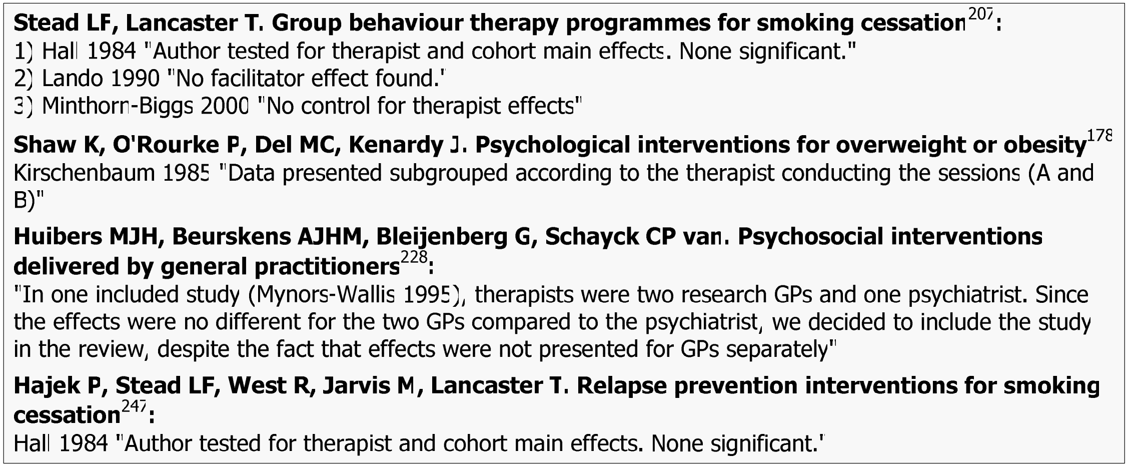
Level of Reporting, n (% total)	**Source of Clustering**		
	Cluster Randomisation	**Group-Based Intervention**	**Therapist Variation**
Cluster Size			
Complete: Explicit	6 (15)	7 (8)	0 (0)
Complete: Via Number of Clusters	4 (10)	0 (0)	3 (3)
Partial: Explicit	3 (8)	12 (14)	2 (2)
Partial: Explicit & Via Number of Clusters	5 (13)	3 (4)	2 (2)
Partial: Via Number of Clusters	7 (18)	1 (1)	23 (23)
Not reported	14 (36)	60 (72)	71 (70)
Intraclass Correlation (ICC)			
Partial: Explicit	3 (8)	0 (0)	0 (0)
Partial: Explicit & Qualitative	1 (3)	0 (0)	0 (0)
Partial: Qualitative	1 (3)	0 (0)	0 (0)
Unavailable but Assumptions Explicit	4 (10)	0 (0)	0 (0)
Not reported	30 (77)	83 (100)	101 (100)
TOTAL	**39**	**83**	**101**

Note: Study ICC estimates could be reported, assumed or reported as completely or specifically unavailable to be counted above

It was apparent that cluster size was viewed as a more general aspect of the study description than as a component of its design effect. Cluster sizes were reported or deducible in a higher proportion of the reviews than ICCs were for all sources of clustering. The number of therapists and the group size tended to be reported rather than the size of therapist caseloads or the group number. In some cases, a range of cluster sizes were reported, although primarily for group-based designs. In others, average cluster sizes were deducible for each study arm. ICCs were either provided explicitly or a qualitative statement was made about their size. They were reported in nine reviews[196, 206, 223, 224, 227, 238, 244, 247, 268], five of these relating to smoking cessation, the best example being that of Carr and Ebbert[227]. None were associated with therapist or group variation. The information that was reported is indicative of what is available in the study reports. While it is likely that ICC estimates were not generally available, only four reviews stated that this was the case for all, or specific, studies[196, 206, 224, 244]. An attempt to contact authors was reported only by Pharoah *et al*[206], and again only in relation to clustering associated with cluster-randomisation.

Handling of potential clustering effects associated with therapist and group variation was not systematically assessed or reported at the study-level in any of the reviews (but see Box 3.4 for sporadic reports). This compared, respectively, to 10 (43%) and 1 (6%) reviews explicitly or otherwise reporting the inclusion of cluster-randomised trials[174, 203, 206, 210, 224, 227, 238, 240, 244, 260, 268]. Even among these reviews, the level of reporting varied with Ebrahim *et al*[260] stating only that the units of randomisation and analysis differed, without then documenting whether allowance had been made for clustering, how, or whether it was sufficient or even appropriate. Similarly, it was not always clear that reviewers understood that clustered studies could be appropriately analysed at the individual level (e.g. Brunner *et al*[203]). It is apparent from Box 3.4 that the use of preliminary tests for the presence of therapist or cohort effects was the only statistical method reported in this context. These tests presumably treated the therapists and groups as fixed-effects, but further relevant information, including *F* ratios, degrees of freedom and *p*-values, were not provided. Neither was comment made on the low statistical power of such tests, their appropriateness or the likely impact absence of adjustment might have on specific study results or conclusions.

Box 3.4 Reporting of Study-Level Handling of Therapist and Group Variation



Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:
1) Hall 1984 "Author tested for therapist and cohort main effects. None significant."
2) Lando 1990 "No facilitator effect found."
3) Minthorn-Biggs 2000 "No control for therapist effects"

Shaw K, O'Rourke P, Del MC, Kenardy J. Psychological interventions for overweight or obesity[178]
Kirschenbaum 1985 "Data presented subgrouped according to the therapist conducting the sessions (A and B)"

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck CP van. Psychosocial interventions delivered by general practitioners[228]:
"In one included study (Mynors-Wallis 1995), therapists were two research GPs and one psychiatrist. Since the effects were no different for the two GPs compared to the psychiatrist, we decided to include the study in the review, despite the fact that effects were not presented for GPs separately"

Hajek P, Stead LF, West R, Jarvis M, Lancaster T. Relapse prevention interventions for smoking cessation[247]:
Hall 1984 "Author tested for therapist and cohort main effects. None significant."

3.4.3 Handling of Study Clustering Effects within Meta-Analyses

Reviewers' handling of within-study clustering effects is summarised in Table 3.9. It is evident that awareness of the clustering implications generalised only beyond the specific scenarios outlined in the Cochrane Handbook[169] in one review[207] (see Box 3.5) and then only in relation to group variation. The typical use of two or three clusters per arm and the common failure to allow for clustering in study analyses was noted. However, reviews considered removing studies within a sensitivity analysis rather than correcting the original analyses at the study-level prior to pooling them. It is not clear what would be gained from a comparison of two meta-analyses, one including and one excluding studies unadjusted for clustering effects. Changes found to the pooled treatment effect, or to the width of the associated confidence interval, would not be attributable to the impact of clustering in these circumstances. The study sample size and its treatment effect would influence any changes, irrespective of the size of the clustering effect, making interpretation of such changes misleading.

Table 3.9 Handling of Clustering Effects in Cochrane Reviews and Meta-Analyses

	Focus of Systematic Review		Overall
Study Design, n (% total)	**Psychotherapy**	**Other**	
Therapist-Delivered Interventions	**0 (0)**	**0 (0)**	**0 (0)**
Group-Based Interventions	**1 (2)**	**0 (0)**	**1 (1)**
Sensitivity analysis considered	1 (2)	0 (0)	1 (1)
Cluster-Randomisation	**20 (34)**	**9 (21)**	**29 (29)**
Studies excluded	2 (3)	1 (2)	3 (3)
Untailored standard paragraph only	6 (10)	1 (2)	7 (7)
Tailored standard paragraph only	1 (2)	0 (0)	1 (1)
Narrative reporting	2 (3)	2 (5)	4 (4)
Subgroup analysis performed	0 (0)	2 (5)	2 (2)
Sensitivity analyses excluding studies	3 (5)	1 (2)	4 (4)
Weights based on number of clusters	1 (2)	0 (0)	1 (1)
Sensitivity analyses including studies	2 (3)	1 (2)	3 (3)
Generic inverse variance method only	1 (2)	1 (2)	2 (2)
Sensitivity analysis based on largest or Range of plausible clustering effects	2 (3)	0 (0)	2 (2)
Crossover	**12 (20)**	**10 (24)**	**22 (22)**
Studies excluded	4 (7)	3 (7)	7 (7)
Narrative reporting	0 (0)	2 (5)	2 (2)
First period data used only	6 (10)	4 (10)	10 (10)
Periods combined if minimal carryover	1 (2)	0 (0)	1 (1)
First period data & sensitivity analysis excluding other crossover trials	0 (0)	1 (2)	1 (1)
Elbourne (2002) methods	1 (1)	0 (0)	1 (1)
TOTAL	**59 (100)**	**42 (100)**	**101 (100)**

While the actions taken by reviewers in relation to cluster-randomised studies reflect awareness of the impact of study design on meta-analyses, complete understanding was not always obvious. Seven (24%) reviews simply included a standard paragraph in the *Methods* section[184, 201, 204, 208, 212, 222, 242] (see Box 3.6), with six of these edited by the Schizophrenia group and only one including relevant studies[222].

Box 3.5 Recognition of the Presence and Implications of Group Variation

PRESENCE

Barlow J, Coren E, Stewart-Brown SSB. Parent-training programmes for improving maternal psychosocial health[216]:
Conclusion "Nixon 1993 also refers to the group process, and the validation of parental feelings through the experience of sharing them with other parents. He also discusses the way in which the process of sharing feelings with other parents contributes to the normalisation and destigmatisation of such feelings, and of the potential for a reduction of negative attributions through comparison with other parents. The extent to which these processes influenced some of the outcomes is discussed, as is the lack of validated measures to assess outcomes of this nature. It is hypothesised that 'social comparison' may well be a powerful treatment tool. Gammon 1991 cites a number of benefits for parents arising from the use of groups in particular, including the possibility for mutual support, the opportunity for learning from each other and for a reduction in feelings of isolation...Research is also needed which focuses on the process of programme delivery. Only four of the primary studies included in this review made any reference to this issue. Of the programme service providers who were contacted, six referred to group factors such as empowerment, support, and self-esteem as being important. Two referred to the role of facilitators in promoting positive group processes."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Discussion "The results of two of the cognitive behavioural therapy trials look encouraging, although with the group-based study we are unable to exclude the possibility that it is the group environment that is having the effect rather than the therapy per se."

IMPLICATIONS

Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:
Methodological Quality "In cases where more than one group method was being compared, and recruitment was continuous, participants were generally allocated to treatment groups on the basis of their sequence of arrival. The group was then randomized to treatment. In studies in which randomization was individual, randomization schedules were in some cases reported to be interrupted in order to allocate families or friends to the same group. Both these features mean that people in a particular group may be more similar than would be expected by chance. This undermines the statistical assumption used to estimate the variance, which is that they are typical of the population as a whole. The same principle also applies when patients are treated in groups, because each person's chance of success may be influenced by the group in which they find themselves. The possibility that success rates varied beyond chance between the groups given the same treatment can be tested, but the power to detect these differences will generally be very low. All these features of group therapy trials are likely to lead to an underestimate of the true variance, and therefore to the estimation of confidence intervals which are too narrow. In those trials which randomized entire worksites to programme type this factor is even more relevant. The small number of trials in any comparison and the fact that studies of the same type tend to share the same shortcomings meant that sensitivity analyses based on any quality assessment were impractical." *Discussion* "A limitation of research in which participants are treated in groups is that typically there may be only two or three groups in each treatment condition. Participants' chances of success are almost certainly not completely independent. There may be variation by the group in which they were treated, due to aspects of the group process. This aspect is generally ignored in trial analyses."

Box 3.6 Standard Paragraphs on Handling of Cluster-Randomisation

VERSION ONE

Methods "Cluster trials: Studies increasingly employ 'cluster randomisation' (such as randomisation by clinician or practice) but analysis and pooling of clustered data poses problems. Firstly, authors often fail to account for intra class correlation in clustered studies, leading to a 'unit of analysis' error (Divine 1992), whereby p values are spuriously low, confidence intervals unduly narrow and statistical significance overestimated causing type I errors (Bland 1997, Gulliford 1999). Secondly, RevMan does not currently support meta-analytic pooling of clustered dichotomous data, even when these are correctly analysed by the authors of primary studies, since the 'design effect' (a statistical correction for clustering) cannot be incorporated. Where clustering was not accounted for in primary studies, we presented data in a table, with an asterisk (*) symbol to indicate the presence of a probable unit of analysis error. In subsequent versions of this review we will seek to contact first authors of studies, to seek intra-class correlation co-efficients of their clustered data and to adjust for this using accepted methods (Gulliford 1999). Where clustering has been incorporated into the analysis of primary studies, then we presented these data in a table. No further secondary analysis (including meta-analytic pooling) will be attempted until there is consensus on the best methods of doing so, and until RevMan, or any other software, allows this. A Cochrane Statistical Methods Workgroup is currently addressing this issue. In the interim, individual studies were very crudely classified as positive or negative, according to whether a statistically significant result ($p<0.05$) was obtained for the outcome in question, using an analytic method which allowed for clustering."

Box 3.6 Continued…

VERSION TWO

Methods "Cluster trials: Studies increasingly employ 'cluster randomisation' (such as randomisation by clinician or practice) but analysis and pooling of clustered data poses problems. Firstly, authors often fail to account for intra class correlation in clustered studies, leading to a 'unit of analysis' error (Divine 1992) whereby p values are spuriously low, confidence intervals unduly narrow and statistical significance overestimated. This causes type I errors (Bland 1997, Gulliford 1999). Where clustering was not accounted for in primary studies, we presented the data in a table, with a (*) symbol to indicate the presence of a probable unit of analysis error. In subsequent versions of this review we will seek to contact first authors of studies to obtain intra-class correlation coefficients of their clustered data and to adjust for this using accepted methods (Gulliford 1999). Where clustering has been incorporated into the analysis of primary studies, we will also present these data as if from a non-cluster randomised study, but adjusted for the clustering effect. We have sought statistical advice and have been advised that the binary data as presented in a report should be divided by a 'design effect'. This is calculated using the mean number of participants per cluster (m) and the intraclass correlation co-efficient (ICC) [Design effect = 1+(m-1)*ICC] (Donner 2002). If the ICC was not reported it was assumed to be 0.1 (Ukoumunne 1999). If cluster studies had been appropriately analysed taking into account intra-class correlation coefficients and relevant data documented in the report, synthesis with other studies would have been possible using the generic inverse variance technique."

Reporting of cluster-randomisation was implicit in this review, which may account for the methods stated not then being implemented. One review[215] tailored the standard paragraph to reflect their recognition that the methods described were theoretical, as no relevant studies were included. Since three reviews excluded cluster-randomised studies[197, 213, 264], this left 19 (49%) of the 39 reviews which included them describing methods for handling associated within-study clustering.

There was some evidence that the specific method reported changed over time, with all four cases of narrative reporting[174, 218, 240, 258] located in reviews not substantially updated since May 2005. Forbes *et al* (1999)[240] intended to report relative risks and confidence intervals in the text, but found that only one study had adjusted for clustering, and published data were unsuitable for calculating relative risks. Rose *et al* (2002)[218] planned to include studies in meta-analyses when guidance was available, and the methods had been implemented in RevMan software. Glasscoe and Quittner (2003)[174] calculated standard errors from adjusted confidence intervals available in the original paper, but reported binary outcomes in the text due to software limitations. Ekeland *et al* (2004)[258] simply stated that cluster-randomised studies were reported in the text due to their design.

The methods described in the remaining 14 reviews can be categorised by how well they addressed the precision implications. Two reviews pooled subgroups of studies defined by their method of randomisation[238, 260] and analysis[260] but did not state how, or whether, allowance was made for clustering in the standard errors. Four reviews excluded cluster-randomised studies in a sensitivity analysis[207, 223, 225, 261], comparing the pooled treatment effects[207, 225, 261] or their statistical significance[223] without making allowance for clustering in either analysis. Both methods accordingly fail to tackle the precision issue. Of these six reviews, only two[223, 260] had been substantially updated since May 2005. Five were edited by the Tobacco Addiction group.

Brunner *et al* (2005)[203], in contrast, over-tackled the issue replacing the number of patients with the number of clusters when calculating study weights, even where published analyses adjusted for clustering and the standard errors were derived from these. The final seven reviews[196, 206, 210, 224, 227, 244, 247] used the generic inverse variance method described in the Cochrane Handbook[169] to allow for clustering. Three compared primary analyses omitting studies to sensitivity analyses adjusting for clustering[206, 210, 247]. One adjusted primary analyses using the reported ICC estimates then performed a subgroup analysis on the basis of the method of randomisation[227]. Another assumed an ICC of 0.1 as realistic for all cluster-randomised studies, and adjusted the primary analyses on this basis[244]. And two performed sensitivity analyses assuming the largest plausible[196, 224] ICC and a range[224] of feasible ICC estimates based on Ukoumunne *et al* (1999)[271], comparing the *p*-values of the pooled treatment effects. Five[196, 203, 206, 227, 244] of these eight reviews had been substantially updated since May 2005.

The level of recognition of the implications for crossover trials was comparable. Ten reviews[171, 172, 180, 184, 185, 220, 250, 252, 255, 267], spanning six Cochrane Review groups, simply stated that data from only the first treatment period would be eligible or analysed in the *Eligibility* or *Methods* section. Crossover trials were identified in only three of these[220, 250, 267]. As a further seven reviews from five Cochrane Review groups excluded crossover trials[176, 179, 182, 218, 236, 239, 263], in two cases because the design was considered inappropriate[179, 182], this left 8 (67%) of the 12 reviews including them reporting methods for handling within-patient clustering. The published data were viewed as insufficient in both cases of narrative reporting[231, 265]. Adjusted confidence intervals were available in one[231], but washout periods were not reported, and the reviewers cited software limitations. The correlation in patient outcomes between periods required for the methods described by Elbourne *et al*[156] was not available in the other[265]. Brunner *et al*[203] intended to pool periods, where the design ensured minimal carryover, but did not state whether or how this applied to the cluster-randomised crossover study they included. Glazener *et al*[269] used first period data where available, and then performed sensitivity analyses comparing the significance of meta-analyses including the remaining crossover trials, as if they had a parallel-group design, to the primary analyses excluding them. Only one review[214] reported utilising the methods described by Elbourne *et al*[156].

3.5 Use and Relevance of Quality Assessment Tools

All Cochrane reviews contain a *Methodological Quality* section in which reviewers are intended to summarise the internal and external validity of included studies along with any variability found in these aspects between studies[169]. Quality assessment is considered essential to "limit bias in conducting the systematic review, gain insight into potential comparisons, and guide interpretation of findings" (p.79)[169]. While the use of scores derived from scales or checklists is not advocated[169], assessment of the potential for selection, performance, attrition and detection biases is. Fifty-two of the 101 reviews explicitly stated using the internal validity criteria recommended in the Cochrane Handbook when assessing study quality. However, the version cited varied from Mulrow and Oxman (1996)[272] through to Higgins and Green (2005)[169], with only 10 of the reviews[196, 201, 206, 212, 222, 230, 234, 242, 249, 263] specifically referring to the latter, all being updated since May 2005, although one not substantially[249].

Box 3.7 Example Comments on the Use of Published Quality Assessment Tools

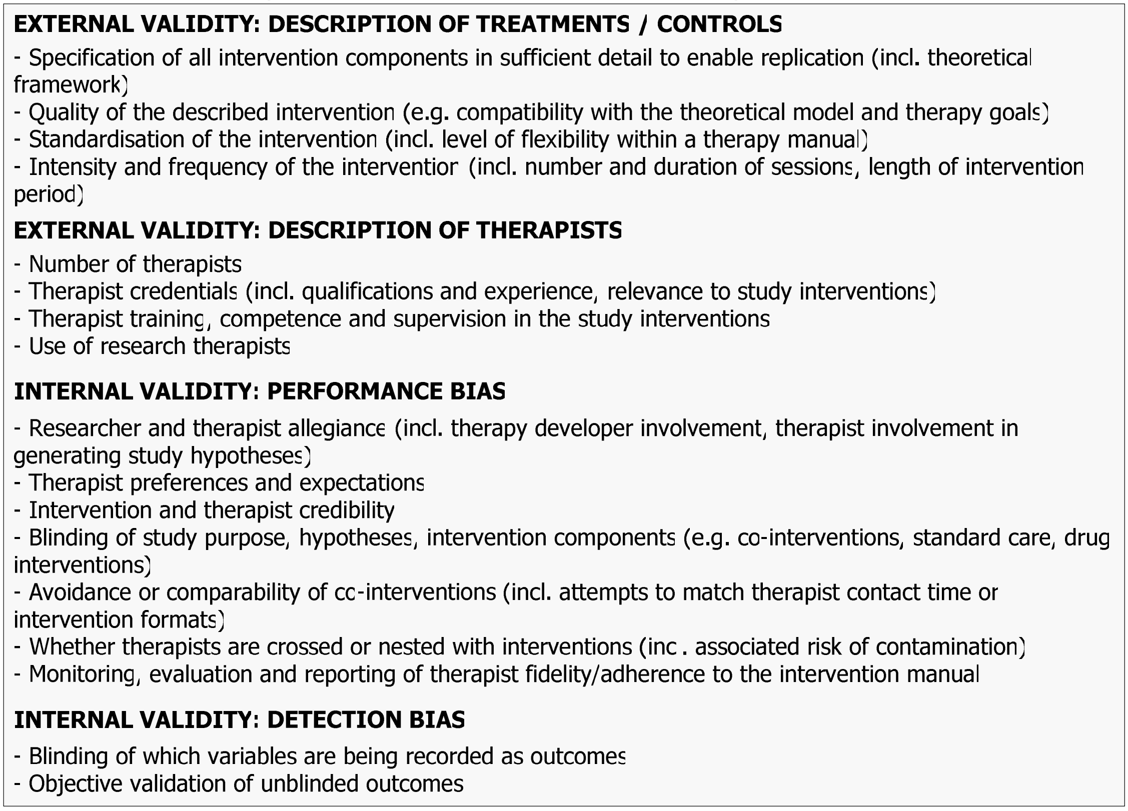
Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders[221]:
Methodological Quality "Some of the elements of the CCDAN scale were not relevant to this type of treatment research. There was no blinding of psychotherapy subjects and specific "side effects" were reported." *Discussion* "The studies were of variable quality...Manuals and adherence measures were not employed in each study calling into question the quality of psychotherapy provided. Therapist experience was in question in many studies, raising the chance that the therapy was not provided in an optimal fashion...The CCDAN Quality Rating System we used did not include ratings on these parameters, which were relevant to the interpretation of psychotherapy study quality"

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Methodological Quality "None of the studies was double blind. It is rarely possible to blind participants or therapists to psychological interventions, so the application of trial quality assessment tools, such as the Oxford scale (Jadad 1996) nor other widely used quality scales (Juni 2001), was not appropriate."

Ostelo RWJG, van-Tulder MW, Vlaeyen JWS, Linton SJ, Morley SJ, Assendelft WJJ. Behavioural treatment for chronic low-back pain[195]:
Methods "The methodological quality of the RCTs was independently assessed...using a criteria list (Table 01) recommended by the Cochrane Back Review Group (van Tulder 1997b; van Tulder 2003)...As it is difficult to blind patients for behavioural treatment, we redefined the criterion regarding the blinding of patients. If blinding was not feasible, item 4 of the criteria list was scored positive if the credibility of the treatments was evaluated and treatments were equally credible and acceptable to patients (Turk 1993)."

The other frequently cited criteria were those of Jadad *et al*[273] used in 15 of the reviews. When adopting these approaches, reviewers often paid particular attention to allocation concealment citing Schulz *et al*[274] and Juni *et al*[275] in support of doing so. The use of standard Cochrane Review group criteria was apparent within the Back[276, 277], Depression, Anxiety and Neurosis[278], Effective Practice and Organisation of Care, and Incontinence[279] groups, with specific scales developed and recommended by the former two. Sporadic use was found[234, 240] of the UK NHS Centre for Reviews and Dissemination guidance[280, 281]. Use of Downs and Black[282] and scales derived from Kenardy and Carr[283] and previous reviews[284, 285] was reported[175, 218, 235, 244]. However, none of these tools explicitly address issues arising from therapist or group variation, neither were the more general items extended to do so. Comments were made on the use of quality assessment tools for studies involving psychotherapy, with additional items being used in some reviews (see Boxes 3.7 and 3.8 respectively for examples).

Box 3.8 Additional Quality Criteria for Studies Involving Psychotherapy



EXTERNAL VALIDITY: DESCRIPTION OF TREATMENTS / CONTROLS

- Specification of all intervention components in sufficient detail to enable replication (incl. theoretical framework)
- Quality of the described intervention (e.g. compatibility with the theoretical model and therapy goals)
- Standardisation of the intervention (incl. level of flexibility within a therapy manual)
- Intensity and frequency of the intervention (incl. number and duration of sessions, length of intervention period)

EXTERNAL VALIDITY: DESCRIPTION OF THERAPISTS

- Number of therapists
- Therapist credentials (incl. qualifications and experience, relevance to study interventions)
- Therapist training, competence and supervision in the study interventions
- Use of research therapists

INTERNAL VALIDITY: PERFORMANCE BIAS

- Researcher and therapist allegiance (incl. therapy developer involvement, therapist involvement in generating study hypotheses)
- Therapist preferences and expectations
- Intervention and therapist credibility
- Blinding of study purpose, hypotheses, intervention components (e.g. co-interventions, standard care, drug interventions)
- Avoidance or comparability of co-interventions (incl. attempts to match therapist contact time or intervention formats)
- Whether therapists are crossed or nested with interventions (incl. associated risk of contamination)
- Monitoring, evaluation and reporting of therapist fidelity/adherence to the intervention manual

INTERNAL VALIDITY: DETECTION BIAS

- Blinding of which variables are being recorded as outcomes
- Objective validation of unblinded outcomes

Based on the following reviews: [171, 172, 174, 180, 183, 185, 192, 194, 195, 198, 213, 218-222, 227, 228, 239, 260, 264]

Blinding of non-pharmacological interventions was seldom considered possible[173, 183, 185, 190, 195, 201, 203, 208, 217, 221, 222, 224, 226-228, 230, 233, 234, 242, 249, 254, 256, 258-261, 264, 266, 270]. This was used simply as a justification for excluding blinding as a quality criterion by some reviewers. Others questioned the relevance or appropriateness of published quality assessment tools on this basis (see Box 3.7). A subset suggested alternative criteria for assessing the extent and impact of performance biases (see Box 3.8), the most common of these being the evaluation of therapist adherence or fidelity. While issues relating to the lack of blinding received most attention, a range of additional criteria were discussed, with exemplars in Abbass *et al*[221], Buckley and Pettit[222], Eccleston *et al*[183], Hajek and Stead[192], Huibers *et al*[228], Hunot *et al*[185] and Littell *et al*[213].

3.6 Recognition of the Implications for Internal Validity

Possible confounding of treatments and therapist characteristics was discussed in six of the reviews (see Box 3.9). Their titles indicate that five were primarily interested in comparisons of different therapeutic approaches, with the other emphasising the comparison of therapist characteristics and packages of the two. As such, imbalance in the characteristics of the therapists across the arms within studies was an issue for internal validity in five[183, 192, 207, 210, 222] with disparities in the therapeutic approaches used threatening internal validity in the other[224]. Recognition of this is clear in four of the quotes (Box 3.9), the two exceptions being those of Lancaster and Stead[210] and Stead and Lancaster[207], which imply awareness of the potential for confounding across but not within studies. The relationship between interventions and therapists, whether it be nested or crossed, was highlighted in three of the reviews[183, 192, 222].

Box 3.9 Confounding of Treatments and Therapist Characteristics

Buckley LA, Pettit T. Supportive therapy for schizophrenia[222]:
Discussion "In trials of psychological interventions, there is question about whether the same therapists should provide two different interventions, potentially allowing factors such as level of experience of the therapists and individual differences in personality to be evenly distributed between groups. However, an alternative, and in our opinion more persuasive, argument is that different therapists should provide different therapies. This takes into account the likelihood that therapists have a loyalty to, and training and experience in, one particular type of therapy. This may be particularly important if they have been involved in generating the hypotheses which are being tested. The majority of studies in this review used the same therapists for supportive therapy and other psychological interventions. Most studies did not specify what training the therapists had received or what level of experience they had. Where studies did report details of therapist training, therapists who delivered supportive therapy were sometimes trained in other modalities, such as cognitive behavioural therapy, but not in supportive therapy. The studies which described standardised supportive therapy, with use of a manual, were in the minority; as were the studies which evaluated or monitored adherence to the treatment model. Not all studies attempted to match the amount of therapist contact. For example, Coyle 1988 compared supportive therapy with social skills training and psychoeducation. Supportive therapy sessions, however, were half as long as other therapy sessions, and were delivered individually rather than in a group."

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Methodological Quality "Therapist / treatment interaction: It is rarely possible to blind participants or therapists in psychological treatments to the treatment offered. Ideally, equivalence of treatment is examined using treatment credibility ratings for participants (as in Griffiths 1996; Larsson 1987a; Larsson 1987b, McGrath 1988; and Sanders 1994). Where therapists are trained in the trial to deliver treatments which are novel to them, credibility ratings are also appropriate. For example, in the trial by Passchier 1990 school staff were trained to deliver two treatments which they found equally credible. Where therapists are already trained in particular treatment techniques, there are two related issues. One is treatment-therapist confounding: if the same therapist gives an active treatment and a placebo control treatment, s/he is

Box 3.9 Continued…

unlikely to believe both to be equally likely to produce improvement. Indicators of therapist allegiance, not reported in any of the trials in this review, may be used to examine this. It is more common, however, for different therapists to be assigned to the treatment conditions according to their skills, with the result that differences between therapists appear as differences between treatments. For six of the trials this did not apply, since the control condition was a waiting list (Barry 1997; Fentress 1986; Labbe 1984; Labbe 1995; Osterhaus 1997; Sanders 1989). For nine studies, therapists were specific to treatment or reporting was unclear. Three studies addressed the issue: McGrath 1988 used three therapists equally and Larsson 1990 used one across treatment conditions; Larsson 1987a used two therapists for both conditions and one for the control only."

Hajek P, Stead LF. Aversive smoking for smoking cessation[192]:
Methodological Quality "Studies in which different therapists run different conditions may be comparing the efficacy of the therapists rather than the efficacy of the methods. Even where the same therapist runs different treatments, the fact that the therapist is not blind and usually believes that one treatment is superior to others can introduce a 'performance bias'. The better studies try to tackle this problem by having several therapists, each running all treatments...Only one of the studies in this review (Hall 1984a) avoids the most glaring methodological problems. All the others present most or all of the following problems: validation not done or incomplete, outcome assessor not blind to subject allocation, different therapists for different treatments or only one therapist involved, no information on continuous abstinence, and very small sample sizes (usually around 20 subjects per condition). Most of these methodological shortcomings can be expected to influence the results in favour of the treatment's efficacy...The poor methodological quality of this body of literature is explained by its age. The methodology of research in smoking cessation has developed considerably over the last 10 to 15 years. Most aversive treatment studies are over 20 years old." *Discussion* "These statistical results must be interpreted in the light of methodological considerations before drawing final conclusions." *Summary* "The results of the existing trials suggest that this may be effective, but the evidence is not conclusive because most of the studies of this approach have methodological problems."

Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation[210]:
Background "One problem in assessing the value of individual counselling is that of confounding with other interventions. For example, counselling delivered by a physician in the context of a clinical encounter may have different effects from that provided by a non-clinical counsellor. One approach to this problem is to employ statistical modelling (logistic regression) to control for possible confounders, an approach used by the US Public Health Service in preparing clinical practice guidelines (AHCPR 1996; AHRQ 2000). An alternative approach is to review only unconfounded interventions. This is the approach we have adopted in the Cochrane Tobacco Addiction Review Group. In this review, we therefore specifically exclude counselling provided by doctors or nurses during the routine clinical care of the patient, and focus on smoking cessation counselling delivered by specialist counsellors."

Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:
Discussion "The US Public Health Service Guideline, Treating Tobacco Use & Dependence (Fiore 2000)...authors stress that the strength of evidence underlying recommendations...is not of the highest level because of the correlation of the types of counselling and behavioural therapies with other treatment characteristics such as programme length or type of therapist. The conclusions of this Cochrane review are consistent with the Guideline finding in relation to the inclusion of general problem-solving components, and are strengthened by being limited to unconfounded comparisons."

Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol[224]:
Description of Studies "The service delivery methods also differed between the studies. Those participants seen by a doctor tended to have less frequent appointments or less time at appointments than those seen by a dietitian (Caggiula 1996; Gosselin 1996; Luepker 1978; Smith 1976)." *Discussion* "It is not possible to distinguish whether the difference in blood cholesterol was a result of advice from a different health professional or from more contact with the health professional. Participants randomised to the dietitian generally received more time with a health professional than those randomised to the doctor."

While all argued that the choice of design was important, there was disagreement about which design was considered superior. Also, while strongly alluded to, a clear distinction was not made, following Wilkins[140], between the function of stable and situation-dependent therapist characteristics when considering one design over the other. As a result, the studies were assessed largely in terms of which design they adopted rather than by whether the design was consistent with their research question or with the associated risk of bias. Buckley and Pettit's[222] concern was chiefly with treatment-dependent factors, leading them to favour a nested design. In contrast, Eccleston *et al*[183] and Hajek and Stead[192] put greater emphasis on therapists' general skill level, and therefore preferred a crossed design.

Imbalance in therapist attributes and behaviours between arms could pose a threat to internal validity, leading to bias, at any stage in the conduct of a study[286]. The quotes in Box 3.9 all relate to therapist-level performance biases. A number of reviewers also considered the possibility of detection biases, with twelve[184, 190, 192, 198, 201, 204, 208, 212, 214, 215, 222, 242] explicitly connecting them to therapist involvement, and two[186, 213] stating that therapists performed outcome assessments. Marshall *et al*[287] was frequently quoted, in reviews edited by the Schizophrenia group, when excluding outcomes assessed by the therapists. The general problem of rater variation was briefly outlined in Glasscoe and Quittner[174], Hajek and Stead[192] and Uman *et al*[181]. For example,

> "There were severe limitations to the method of scoring which involved interviewing skills and subjective judgement on the part of the raters with consequent differences in individual style of delivery. There was an acknowledgement that these differences…may have influenced the results." Glasscoe and Quittner[174] (pp.12-13)

This has many parallels with therapist and group variation, and indeed these sources of variation may even be aliased in some circumstances[286].

One implication of therapist and group-based intervention studies is that they entail multiple allocations of treatments to experimental units[286]. For instance, the simplest nested therapist designs necessitate allocation not only of interventions to patients but of interventions to therapists and of therapists to patients. Each of these allocations may be random or non-random, concealed or unconcealed, with selection bias arising potentially from any non-random or unconcealed allocation. The impact this has on the validity of interpretations at the study-level will depend on the match between the research question and the chosen design[286]. It is influenced in a systematic review by the compatibility between the designs of included studies and the research question of the review.

Instances of multiple randomisations found in the 101 reviews are given in Box 3.10. These were found in the *Characteristics of Included Studies* table in the appendix of each review unless stated otherwise. None of the instances reported were associated with assignment of groups in group-based intervention studies. Hunot *et al*[185] was the only review to report the presence of multiple randomisations systematically. Given the emphasis on single randomisations, individual- or cluster-, it is likely that other instances went unreported. Hunot *et al*[185] did not report what was randomly allocated to what; this had to be confirmed by referring back to the original papers. None of the reviews recorded whether the additional allocations were concealed, or the methods reported for doing so. There was no discussion of the consequences of this aspect of the study design for the interpretation of the results either.

Box 3.10 Reporting of Multiple Randomisations

MULTIPLE RANDOMISATION: INTERVENTIONS TO PATIENTS & INTERVENTIONS TO THERAPISTS

Hajek P, Stead LF, West R, Jarvis M, Lancaster T. Relapse prevention interventions for smoking cessation[247]:
1) Hall 1984 "Randomization: method NS...Therapists: 2 psychologists, randomly assigned to groups"
2) Killen 1984 "Randomization: method NS (married couples allocated to same condition)...Behaviour therapy provided by 2 psychologists, 1 MSW, assigned randomly to treatment conditions"

Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:
1) Gruder 1993 "Randomization: to group or no-group at time of registration. No details on method. 1205 subjects assigned to a group condition, and attempts made to contact them to schedule group meetings. Randomization between the two group conditions was by site...Therapists: Mainly nurses and health educators randomly assigned and trained to lead either Social Support or Discussion meetings."
2) Hall 1984 "Randomization: randomly assigned, no details...Therapists: 2 psychologists, randomly assigned to groups"
3) Killen 1984 "Randomization: Method not stated. individual randomization...Therapists: 3: 2 psychologists, 1 medical social worker, assigned randomly to treatment conditions"

MULTIPLE RANDOMISATION: INTERVENTIONS TO PATIENTS & THERAPISTS TO PATIENTS

Crawford-Walker CJ, King A, Chan S. Distraction techniques for schizophrenia[204]:
Tarrier 1993 "Allocation: randomly allocated to a treatment group and a psychologist."

Hunot V, Churchill R, Silva-de LM, Teixeira V. Psychological therapies for generalised anxiety disorder[185]:
Description of Studies "Design: All the studies included in the review were described as randomised controlled trials, with randomisation at the patient (n=16) or patient and therapist level (n=9)."
1) Arntz 2003 "Allocation: randomised at patient and therapist level - method not reported"
2) Barlow 1992 "Allocation: randomly assigned to treatment condition and to available therapists"
3) Blowers 1987 "Allocation: randomised at patient and therapist level - method not reported"
4) Borkovec 1987 "Allocation: randomised in 3 waves at patient and therapist level - no further information reported"
5) Butler 1991 "Allocation: randomised at patient and therapist level - method not reported"
6) Durham 1987 "Allocation: randomised at patient and therapist level - method not reported"
7) Durham 1994 "Allocation: randomised at patient and therapist level - method not reported"
8) Gath 1986 "Allocation: randomised at patient and therapist level - method not reported"

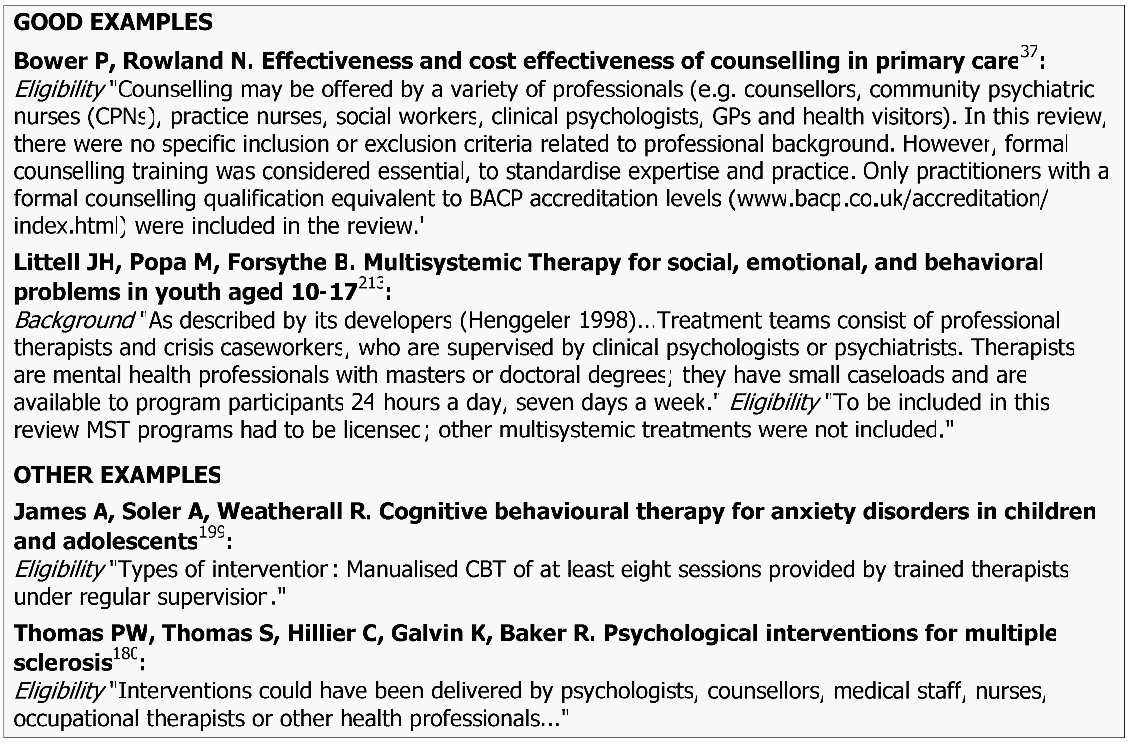
Littell JH, Popa M, Forsythe B. Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17[213]:
Borduin 1995 "Random assignment to treatment conditions and to therapist within conditions."

As there was no recognition of either the precision or selection bias implications due to therapists, it is unsurprising that the impact of withdrawals and losses to follow-up was overlooked at this level either. Barlow *et al*[216] discussed an association between therapist inexperience and patient withdrawals from parent-training programmes, and drew attention to the importance of intention-to-treat analyses for limiting attrition biases. However, they did not consider the contribution of excluding patient outcomes linked with particular therapists or therapist turnover as potential sources of attrition bias. Where aspects of the design are observational, the role of intention-to-treat analyses is less clear. While further methodological work is needed in this area, it is evident that the role of therapist-level attrition also deserves attention.

3.7 Recognition of the Implications for External Validity

Twenty-eight of the 101 reviews referred to therapist characteristics in their eligibility criteria (see Box 3.11 for examples), 22 (79%) of these had a psychotherapy focus. It is of note that therapists' professional background, qualifications and training were the characteristics reviewers mentioned, with supervision arrangements referred to in 5 of the reviews[182, 199, 213, 236, 239]. None included the level, or relevance, of therapists' experience, or whether they were employed as research therapists. The rationale for the choice of criteria was not always clear, as little indication was often given of the therapist population of interest to the review, be it expert or representative of clinical practice. The basis on which criteria were assessed was not always apparent either.

Box 3.11 Examples of Inclusion of Therapist Characteristics in Review Eligibility Criteria



GOOD EXAMPLES

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Eligibility "Counselling may be offered by a variety of professionals (e.g. counsellors, community psychiatric nurses (CPNs), practice nurses, social workers, clinical psychologists, GPs and health visitors). In this review, there were no specific inclusion or exclusion criteria related to professional background. However, formal counselling training was considered essential, to standardise expertise and practice. Only practitioners with a formal counselling qualification equivalent to BACP accreditation levels (www.bacp.co.uk/accreditation/index.html) were included in the review."

Littell JH, Popa M, Forsythe B. Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17[213]:
Background "As described by its developers (Henggeler 1998)...Treatment teams consist of professional therapists and crisis caseworkers, who are supervised by clinical psychologists or psychiatrists. Therapists are mental health professionals with masters or doctoral degrees; they have small caseloads and are available to program participants 24 hours a day, seven days a week." *Eligibility* "To be included in this review MST programs had to be licensed; other multisystemic treatments were not included."

OTHER EXAMPLES

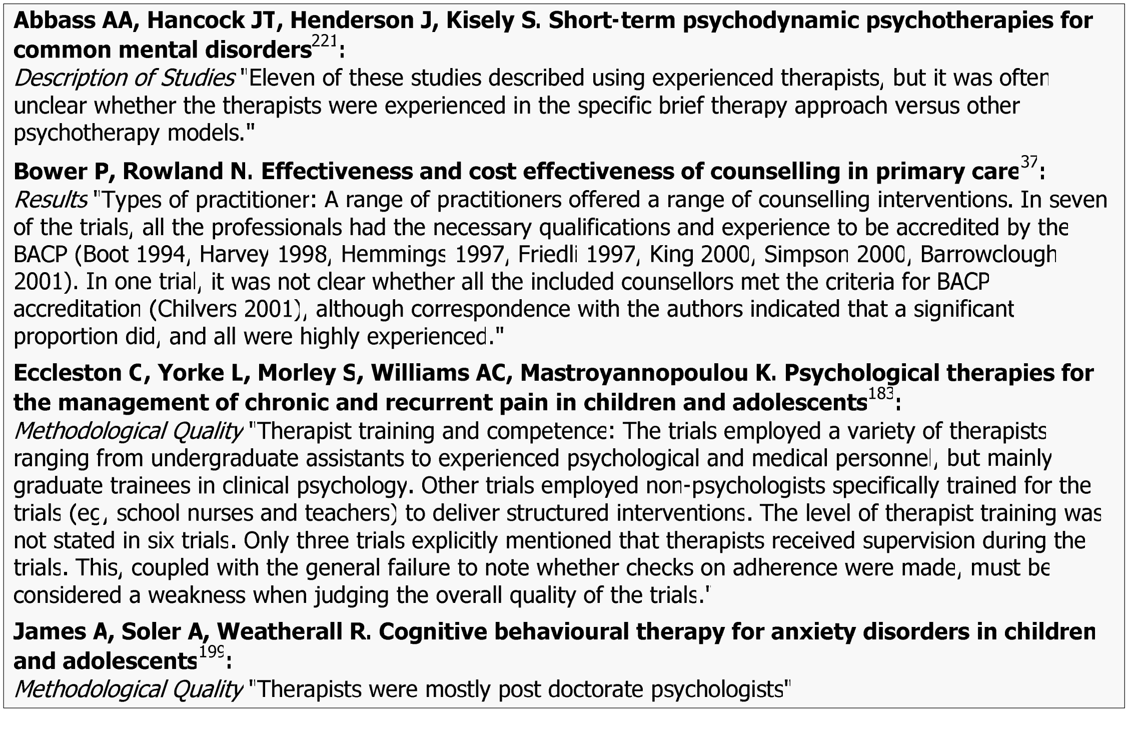
James A, Soler A, Weatherall R. Cognitive behavioural therapy for anxiety disorders in children and adolescents[199]:
Eligibility "Types of intervention: Manualised CBT of at least eight sessions provided by trained therapists under regular supervision."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Eligibility "Interventions could have been delivered by psychologists, counsellors, medical staff, nurses, occupational therapists or other health professionals..."

Some simply stated that therapists had to be trained or qualified, without then giving sufficient detail to specify the therapist population or facilitate replication. Examples of more precise criteria were found in Bower and Rowland[37] and in Littell *et al*[213]. Even here, little or no information was given about the therapists delivering the control, any co-interventions or standard care.

Characteristics of the therapist samples were reported in the text, the *Characteristics of Included Studies* table, or in both of these, for one or more of the included studies involving psychotherapy in 78 (77%) reviews. Examples of statements found in the text are given in Box 3.12. Barbato and D'Avanzo[211] was the only review to refer to therapist characteristics in the eligibility criteria without giving details of the therapist samples included in the studies. James *et al*[199] included a brief description of the therapist sample in the text but did not report study-level characteristics. Of the 77 reviews that did, 24 (31%) also included a statement in the text. These suggest that, while this information was often in the study reports, the descriptions were not always adequate, a pattern of reporting mirrored in the reviews themselves. The statements were mainly located in the *Description of Studies* section[185, 188, 195, 201, 210, 217, 221, 223-227, 229, 236, 243, 245, 248, 249, 251, 259, 268], but were also found in the *Abstract*[259], *Methodological Quality*[183, 199], *Results*[37, 259], *Discussion*[173, 226, 259] and *Conclusions*[173] sections.

Box 3.12 Examples of Reporting of Therapist Samples in the Review Text



Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders[221]:
Description of Studies "Eleven of these studies described using experienced therapists, but it was often unclear whether the therapists were experienced in the specific brief therapy approach versus other psychotherapy models."

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Results "Types of practitioner: A range of practitioners offered a range of counselling interventions. In seven of the trials, all the professionals had the necessary qualifications and experience to be accredited by the BACP (Boot 1994, Harvey 1998, Hemmings 1997, Friedli 1997, King 2000, Simpson 2000, Barrowclough 2001). In one trial, it was not clear whether all the included counsellors met the criteria for BACP accreditation (Chilvers 2001), although correspondence with the authors indicated that a significant proportion did, and all were highly experienced."

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Methodological Quality "Therapist training and competence: The trials employed a variety of therapists ranging from undergraduate assistants to experienced psychological and medical personnel, but mainly graduate trainees in clinical psychology. Other trials employed non-psychologists specifically trained for the trials (eg, school nurses and teachers) to deliver structured interventions. The level of therapist training was not stated in six trials. Only three trials explicitly mentioned that therapists received supervision during the trials. This, coupled with the general failure to note whether checks on adherence were made, must be considered a weakness when judging the overall quality of the trials.'

James A, Soler A, Weatherall R. Cognitive behavioural therapy for anxiety disorders in children and adolescents[199]:
Methodological Quality "Therapists were mostly post doctorate psychologists"

Only 30 of the reviews systematically reported therapist characteristics at the study level (Table 3.10), with reporting for non-experimental interventions being particularly poor. Five reviews reported the number of therapists systematically per study[37, 192, 228, 229, 246]. Where recorded, this ranged from 1 to 156, with most studies involving less than 10 therapists. The level of reporting was not clearly associated with the focus of the review. Reviewers' awareness of the implications for external validity was overtly reflected in their discussion of whether the results could be generalised beyond the therapists involved in the studies. Twenty-one reviews made relevant comments in their *Discussion* or *Conclusions* sections; 10 reviews (48%) referred to therapists in the eligibility criteria[37, 180, 201, 210, 222, 224, 226, 228, 229, 248] and 11 had not[172, 173, 185, 194, 197, 214, 215, 217, 225, 254, 261]. Seventeen reviews (81%) focused on psychotherapy. Examples of the comments are categorised in Box 3.13 by whether the recognition was explicit or implicit.

Table 3.10 Study-Level Reporting of Therapist Number and Characteristics

Level of Reporting	**Focus of Systematic Review**					**Overall**
	Psychotherapy			**Other**		
	General	Specific	Therapist Characteristics	Wider	Different	
Number of Therapists by Study, n (% total)						
Systematic and complete	0 (0)	1 (3)	1 (20)	1 (5)	0 (0)	3 (3)
Systematic but incomplete	0 (0)	1 (3)	0 (0)	1 (5)	0 (0)	2 (2)
Unsystematic	0 (0)	4 (12)	0 (0)	1 (5)	1 (5)	6 (6)
Minimal	7 (33)	4 (12)	1 (20)	2 (9)	3 (14)	17 (19)
Not reported	14 (67)	23 (70)	3 (60)	16 (76)	17 (81)	73 (72)
Therapist Characteristics by Study, n (% total)						
Systematic and complete	2 (9)	2 (6)	3 (60)	7 (33)	2 (10)	16 (16)
Systematic but incomplete	2 (9)	6 (18)	2 (40)	1 (5)	3 (14)	14 (14)
Unsystematic	6 (29)	7 (21)	0 (0)	3 (14)	4 (19)	20 (20)
Minimal	10 (48)	6 (18)	0 (0)	8 (38)	3 (14)	27 (27)
Not reported	1 (5)	12 (36)	0 (0)	2 (10)	9 (43)	24 (24)
TOTAL	**21**	**33**	**5**	**21**	**21**	**101**

Note: This is restricted to studies involving psychotherapy

The limitations associated with studies recruiting highly expert or research therapists were highlighted in 7 of the reviews[173, 185, 194, 197, 217, 222, 228]. Another 4 raised the issue of access to therapists with the required training in clinical practice. None indicated whether their comments applied equally to the interpretation of all the meta-analyses reported, or concluded that external validity had been established.

Box 3.13 Examples of the Discussion of Generalisation of Results to Therapist Populations

EXPLICIT COMMENTS

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Conclusion "The results can only be generalised to similar patients and counsellors. This means that the evidence is restricted to counsellors with BACP accreditation or equivalent.'

Box 3.13: Continued...

Eustice S, Roe B, Paterson J. Prompted voiding for the management of urinary incontinence in adults[217]:
Discussion "External validity has been weakened due to the use of research staff to implement the intervention in four of the trials (Hu 1989; Ouslander 2005; Schnelle 1989; Schnelle 2003). The other trials have demonstrated that it is possible to use direct caregivers during the trial period (Engberg 2002; Linn 1995; Smith 1992; Surdy 1992; Schnelle 1983)."

IMPLICIT COMMENTS

Buckley LA, Pettit T. Supportive therapy for schizophrenia[222]:
Conclusion "Future trials should clearly explain whether practitioners who deliver supportive therapy have been specifically trained, and if so how. It may make the results more applicable if the therapists are trained but in the context of routine career development, rather than specific highly-trained specialised practitioners."

den-Boer PCAM, Wiersma D, Russo S, van-den-Bosch RJ. Paraprofessionals for anxiety and depressive disorders[226]:
Conclusion "Significant questions remain about the conditions under which paraprofessionals can be effective. Most studies mention some selection, training and supervision of paraprofessionals. If paraprofessionals, volunteers or patients, can be effective therapists (with no training or minor initial training), or can offer support because of their personal experience with the underlying problem, this will bring psychological treatment within the scope of psycho-education or education alone."

Joy CB, Adams CE, Rice K. Crisis intervention for people with severe mental illnesses[201]:
Discussion "It is unfortunate that no data are available for staff satisfaction. Issues such as staff recruitment, despondency and burnout are essential to the successful implementation of home care packages. Several of the studies mentioned these as notable problems affecting the running of the project. If such problems were prominent in these usually well-resourced and well-motivated research teams, they may amount to insurmountable obstacles to the implementation of similar projects in routine psychiatric settings."

3.8 Conclusions

It is clear from the quotes extracted in this review that some Cochrane reviewers are aware of the complexities arising from therapist variation and have highlighted those of particular relevance to their research questions. It can be concluded, therefore, that they are aware of the methodological literature published over the past 60 years in subject-specific journals. There was no suggestion, in contrast, that they were familiar with the more recent statistical literature on therapist-related clustering effects[81-84], or that they recognised its implications for assessing study quality or meta-analyses. This is unsurprising because the guidance available in the Cochrane Handbook[169] reflects the emphasis given to specific trial designs and the precision implications in the statistical literature. This was generalised in just one[207] of the 101 reviews, and then only to the groups in studies of group-based interventions. None of them made any reference to care providers, outcome assessors, or clinical services as sources of treatment-related clustering. Likewise none mentioned more efficient methods for incorporating repeated measurements, at baseline[149, 150] or follow-up. The reporting guidelines for systematic reviews and meta-analyses[288] overlook within-study clustering as well. Perhaps it is now time to start bringing the emerging statistical and the wider methodological literatures together.

A number of recommendations could be made on the basis of this review. Firstly, both the multi-component nature of complex interventions and the multilevel nature of their trial designs should be considered throughout; they have implications for the research questions of a systematic review or meta-analysis, its eligibility criteria, data collection, quality assessment, choice of summary measure, methods of synthesis, presentation and interpretation of results. Secondly, given the complexities inherent in the structure of the data, the current absence of accepted statistical methods for handling many of them, and the likelihood that important information won't be easily accessible, it would be worthwhile obtaining the individual-patient-data in future. This would not only give those performing meta-analyses greater flexibility in their choice of approach, but also avoid the need for review conclusions to be conditioned on assumptions about missing data, e.g. ICC estimates or cluster size distributions.

Further research is also needed to address many of the issues raised by these reviews. In particular, statistical methods are required that permit treatment-related clustering effects to be incorporated into meta-analyses. These will differ, to some extent, by the specific multilevel trial designs and their associated data structures. However, as some of these designs are special cases of others there is the scope to develop more general methods. Similarly, although a new scale has recently been published for assessing the quality of psychotherapy trials[289], it was derived solely from the psychotherapy-specific literature. Since all of the issues generalise beyond psychotherapy, combining scales of this kind for all complex interventions and then regularly updating them in light of new developments, might avoid duplication of effort and encourage efficient and timely use of the entire methodological literature.

Although now obvious in retrospect, it became increasingly apparent during the course of this review that the data to be synthesised were qualitative rather than quantitative. While best practice for carrying out systematic reviews was followed as far as possible, it remains an open question as to whether it could be regarded an exemplary example of a systematic narrative review. As the aim was to explore the range and complexity of issues in Cochrane reviews and their recognition by Cochrane reviewers, difficulties encountered when estimating the prevalence of specific study designs is an outcome rather than a limitation. It is reasonable to expect reviewers to report information as unavailable-those reporting characteristics systematically did just this. Further research is required, however, with the study reports and contact with authors as the data, for prevalence estimates to inform the priorities for future methodological research in this currently fast moving field.

3.9 Appendix: Full Quotes taken from Cochrane Reviews

Box 3.14 General Awareness of the Presence of Therapist Variability

Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders[220]:
Abstract "...variability in treatment delivery and treatment quality may limit the reliability of estimates of effect for STPP."

Barlow J, Coren E, Stewart-Brown SSB. Parent-training programmes for improving maternal psychosocial health[215]:
Abstract "Whilst the results of this review are positive overall, some studies showed no effect. Further research is needed to assess which factors contribute to successful outcomes in these programmes with particular attention being paid to the quality of delivery." *Discussion* "Research has shown that failure...to persist through the programme itself, is associated with therapist inexperience (Frankel 1992). These problems surrounding the issue of...drop-outs point to the importance of evaluating the results of trials on an intention-to-treat basis which would limit bias arising from this source."

Buckley LA, Pettit T. Supportive therapy for schizophrenia[221]:
Discussion "Psychotherapy relies on the uniqueness of the clinician-patient relationship, and ways of measuring outcome which take account of this need to be developed (Holmes 2000)."

Ebrahim S, Beswick A, Burke M, Davey SG. Multiple risk factor interventions for primary prevention of coronary heart disease[259]:
Methodological Quality "It is likely that the quality of the intervention, in terms of...person carrying out activities...will determine the impact of intervention."

Eustice S, Roe B, Paterson J. Prompted voiding for the management of urinary incontinence in adults[216]:
Discussion "The nursing home environment and the attitudes of staff are likely to impact on the patients' ability to maintain continence on admission and these are issues that deserve investigation."

Hajek P, Stead LF. Aversive smoking for smoking cessation[190]:
Methodological Quality "it is generally believed that the same method can achieve different results when applied by different therapists."

Hay PJ, Bacaltchuk J, Stefano S. Psychotherapy for bulimia nervosa and binging[189]:
Results "the specialist clinic care in Durand 2003 may have been of variable quality"

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck CP van. Psychosocial interventions delivered by general practitioners[227]:
Discussion "The effectiveness of any intervention is influenced by the triad of the intervention-receiver (e.g. the patient), the intervention-giver (e.g. the GP) and the intervention itself. These three factors are inevitably linked, and any disturbance in this relation at any point can result in negative effects."

Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation[209]:
Discussion "Although we cannot exclude the possibility that small differences in...the therapists' training or skills, have an effect on the outcome, it is not possible to detect such differences in the meta analysis."

Littell JH, Popa M, Forsythe B. Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17[212]:
Background "Considerable attention has been paid to the transportability and dissemination of MST, and to the fidelity of MST replications (e.g., Henggeler 2002b, Schoenwald 2000b, Schoenwald 2001)." *Methods* "The decision to pool results was driven by claims that positive effects of MST are reliable 'across problems, therapists, and settings' (Kazdin 1998) and the practice of combining outcomes across populations and comparison conditions in previous reviews of MST (e.g., Curtis 2004)."

Montgomery P, Dennis J. Cognitive behavioural interventions for sleep problems in adults aged 60+[196]:
Discussion "Homogeneity: Specific cognitive-behavioural interventions and the mode and quality of therapist delivery vary somewhat. It may be that these differences can explain some of the heterogeneity in these results."

O'Connor AM, Stacey D, Entwistle V, Llewellyn TH, Rovner D, Holmes RM, Tait V, Tetroe J, Fiset V, Barry M, Jones J. Decision aids for people facing health treatment or screening decisions[253]:
Discussion "As well, satisfaction could be more strongly affected by the relationship with the practitioner than the decision aid"

Rice VH, Stead LF. Nursing interventions for smoking cessation[224]:
Discussion "In some studies the proposed intervention was not delivered consistently to all participants...Almost all the intensive interventions were delivered by either dedicated project staff or nurses

Box 3.14 Continued...

with a health promotion role. Most studies in which an intensive intervention was intended to be delivered by a nurse with other roles, reported problems in delivering the intervention consistently. None showed a statistically significant benefit of intervention.

Rose S, Bisson J, Churchill R, Wessely S. Psychological debriefing for preventing post traumatic stress disorder (PTSD)[217]:
Discussion "...it is difficult to see how a shame based reaction could be elicited without a skilled, attuned and sensitive therapist. It may however, indicate that a 'safer' way of handling early psychological interventions is to elicit a client led narrative without insisting on a clinician led re-exposure to the event."

Spector A, Orrell M, Davies S, Woods B. Reality orientation for dementia[218]:
Discussion "the entire concept of assessing the success of any psychological therapy can be highly problematic, as it is not possible to account for variables such as the therapeutic alliance between patients and therapists, and the sensitivity with which the therapy is given. It is difficult to assess the more subjective aspects of RO just by reading a written account, yet it may be these very variations which produced variations in results." *Conclusion* "As with all psychological interventions, the success of RO may be dependent on it being used at the appropriate time, by a sensitive and experienced practitioner, to a receptive patient."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[178]:
Background "a number of methodological challenges were evident...The type of intervention, content, theoretical basis, intensity, duration, length of each session, whether one-to-one or in groups can vary, as can the profession and experience of the person delivering the intervention, and the location. This heterogeneity could make it difficult to combine the results from different studies."

Woods B, Spector A, Jones C, Orrell M, Davies S. Reminiscence therapy for dementia[219]:
Background "In the UK the development of the 'Recall' tape-slide package (Help the Aged 1981) meant that reminiscence triggers were widely available in day care centres, care homes and hospitals, leading many staff to establish some form of reminiscence work of variable quality."

Box 3.15 Conventional Explanations for the Consequences of Therapist Variability

TREATMENT STANDARDISATION AND THERAPIST COMPETENCY

Hackett ML, Anderson CS, House AO. Interventions for treating depression after stroke[239]:
Discussion "For psychotherapy trials, there is also good evidence that efficacy is linked to delivery of an adequate exposure to the intervention. This means that therapists should be trained and supervised in the therapy they are delivering, and use a standardised, pre-specified, framework for therapy. To achieve this in psychotherapy trials, the therapy is often manualised and the research therapists are trained and supervised in the use of the manual. Success in brief therapy is linked to adherence to the therapeutic model as well as to the therapists' characteristics. Future stroke psychotherapy trials should also adhere to these standard psychotherapy research guidelines if there is to be any probability of demonstrating consistency and response."

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Eligibility "A distinction can be made between explanatory and pragmatic RCTs. The former seek to impose the highest practical levels of control on variables (e.g. length of treatments, expectancy effects) in order to isolate the key 'active ingredients' and provide a valid test of whether a specific intervention is influencing outcome. Such trials focus on the internal validity of the study. In contrast, pragmatic trials seek to determine the relative 'value' of treatments as they would be provided in routine care settings, and seek to increase external validity without significantly compromising internal validity. In primary care, this means that interventions are not highly standardised, so as to reflect the clinical variation that exists in routine care contexts."

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck CP van. Psychosocial interventions delivered by general practitioners[228]:
Eligibility "Although explanatory studies serve the objective of this review best, pragmatic trials were also eligible for entry in the review as long as the psychosocial intervention was standardised to some degree"

"NON-SPECIFIC" OR PLACEBO EFFECTS

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Discussion "Psychological interventions are complex in that they usually consist of a number of different elements. Some of these elements will be active ingredients specifically included because they are based on psychological theory. Other elements may not be specific to psychological interventions and may be common to many different types of intervention (such as interacting with other people with MS in a group). Still other elements will be specific to individual therapists (for example the therapist's experience and enthusiasm, and the way the therapist interacts with the client)."

Box 3.15 Continued…

Merry S, McDowell H, Hetrick S, Bir J, Muller N. Psychological and/or educational interventions for the prevention of depression in children and adolescents[245]:
Background "Psychological interventions may appear to be effective ”not because of the theories or therapeutic procedures but because of underlying, unspecified or not clearly determined non-specific effects“ (Shapiro 1997 p103). This is of relevance in prevention programs where interventions designed to appeal to participants and introduced by enthusiastic research teams could lead to reduction in depression, at least in the short-term. Improvement in mood may then be attributed to the content of the program. Ideally, as in medication trials, the intervention should be compared with a comparison condition that resembles the intervention but without the elements thought to be actively therapeutic (Shapiro 1997). At the least there should be some attempt to ensure that participants in the study do not know whether they are subjects or controls. If this is not done it is difficult to ensure that effects reported are not placebo effects."

Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents[181]:
Eligibility "nonspecific-treatment or “attention-placebo” control...a group that engages in all of the accouterments of the intervention (e.g., meeting with a therapist, receiving an explanation for the problem) but not the key components of the intervention; used to determine if the effects of the intervention are due to nonspecific treatment components (Kazdin 2003)."

Glazener CMA, Evans JHC, Cheuk DKL. Complementary and miscellaneous interventions for nocturnal enuresis in children[231]:
Discussion "One factor which makes interpretation of trials of complementary treatment problematic is that it is often difficult to disentangle the effect of time spent with a therapist (ie the placebo or psychological support provided) from the actual treatment being tested."

Hajek P, Stead LF. Aversive smoking for smoking cessation[192]:
Eligibility "The task of the review was to see if aversion therapy has a specific effect, i.e. an effect over and above non-specific factors inherent in therapist contact. Comparisons of aversion treatment with no treatment were not included. In most studies there were ′attention placebo′ or other controls roughly matched for therapist contact, although in a few the aversion treatment subjects had up to twice as many treatment sessions as controls."

Abbot NC, Stead LF, White AR, Barnes J. Hypnotherapy for smoking cessation[209]:
Conclusion "Comparison needs to be made with active interventions, preferably matching for therapist contact time."

Morriss RK, Faizal MA, Jones AP, Williamson PR, Bolton C, McCarthy JP. Interventions for helping people recognise early signs of recurrence in bipolar disorder[173]:
Results "Two studies included psychological intervention that did not include EWS component along with TAU (Colom 2003c; Colom2003b) to control for the non-specific effects of psychological treatment, including time spent with therapists."

Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder[214]:
Discussion "In the broader field of psychotherapy research, similar constructions of “placebo” therapy to control for the therapist's attention and the non-specific elements have been broadly used (Kendall 2004, pp. 20-21). However, recent research on the common factors in psychotherapy raise the question of how adequate it is conceptually, and also whether it is technically possible, to separate the active from the non-active elements of therapy (Lambert 2004, pp. 150-152). In any case, the results of the included studies are likely to underestimate the true effects of music therapy, because the control conditions contain a number of potentially efficacious techniques which are also used in music therapy."

Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:
Discussion "...as with all behavioural as opposed to pharmacological therapies, the choice of an appropriate control condition presents problems when evaluating efficacy. There is no obvious equivalent for the drug placebo to control for the non-specific effects of a treatment method. Evaluating group therapies against a waiting list control does not provide very good evidence for the specific effect of the group format."

Buckley LA, Pettit T. Supportive therapy for schizophrenia[222]:
Description of Studies "Thirteen of the twenty-one studies attempted to match experimental and control psychosocial interventions for amount of therapist contact (Eckman 1992, Falloon 1982, Haddock 1999, Kemp 1996, Levine 1998, Lewis 2002b, Pinto 1999, Sensky 2000b, Spaulding 1999, Tarrier 1998, Telles 1995, Turkington 2000, Wirshing 1991). In contrast, four studies took the approach that different interventions by their nature involve different amounts of therapist contact (Dincin 1982, Hogarty 1997-study 1, Hogarty 1997-study 2, Stanton 1984). The other studies did not report on this matter."

“PROCESS” RESEARCH

Barlow J, Coren E, Stewart-Brown SSB. Parent-training programmes for improving maternal psychosocial health[216]:
Conclusion "There is very little research available to date addressing the role of ′process′ factors, such as the way in which the programme is delivered, in producing positive outcomes with regard to parental

Box 3.15 Continued…

functioning. However, it seems likely that the group facilitator/leader has an important part to play in helping parents not only to persist with a particular programme (Frankel 1992), but in facilitating an atmosphere of openness and trust between the participating parents, and in helping parents to feel respected, understood, and supported. Group leaders can play an important role in modelling attributes such as empathy, honesty and respect, and personal qualities such as a sense of humour, enthusiasm, flexibility, and warmth. The absence of data on process factors in the studies that were included in this review precludes the possibility of assessing to what extent the lack of positive change, where this occurred, was due to the content of the programme or its delivery...factors [that] are specific to particular training programmes or non-specific factors such as group-leader qualities...However, as discussed above, the literature lacks any discussion of the process of service delivery or its impact on psychosocial outcomes. Future research would benefit from some consideration of the impact of such factors on the outcomes recorded."

Box 3.16 Therapist Variation as Arm or Study Characteristics

REVIEWS FOCUSING ON THERAPIST CHARACTERISTICS

Carr AB, Ebbert JO. Interventions for tobacco cessation in the dental setting[227]:
Background "Compared to other health care providers, dentists more accurately estimate patient tobacco use (Block 1999). However, dental practitioners are less consistent with and supportive of intervention, less likely to report having strong knowledge or skill levels regarding tobacco cessation, and more likely to perceive barriers to tobacco intervention (Block 1999)." *Methods* "We hypothesized that the following would explain heterogeneity which was explored through subgroup analyses...interventions delivered by dentists versus dental hygienists or other dental staff"

den-Boer PCAM, Wiersma D, Russo S, van-den-Bosch RJ. Paraprofessionals for anxiety and depressive disorders[226]:
Background "The term 'paraprofessional' generally describes a whole category of mental health personnel who are not qualified as psychiatrists, psychologists, social workers or nurses, and who are below a master's-degree level of education (Moffic 1984). Alternatively, paraprofessionals may be experienced patients, residents from local catchment areas (Grant 1996) or college students (Sherman 1998)...All have had some degree of training, are connected to professional staff and supervised by professionals in the work they are doing to ensure quality of care and communication skills, and to prevent emotional burn out. On a number of points, the quality of the relationship with clients may differ between paraprofessionals and professionals. Often paraprofessionals ground their therapeutic relationship not so much in established theory or empirical research but in day-to-day experience and commonsense (Rohde 1996)...If paraprofessionals (like lay people or clients themselves) can perform effective psychological treatment (with or without some initial training, but not to a qualification degree) under (or without) supervision by a professional, then this will bring psychological treatment within the range of psycho-education, or even simply education...This review aims to critically examine the commonsense notion that professional training/ qualification is necessary to deliver effective psychological treatment for anxiety and depressive disorders."

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck CP van. Psychosocial interventions delivered by general practitioners[228]:
Background "the psychosocial interventions addressed in the aforementioned reviews were mostly performed by primary care workers other than the GP (e.g. nurses, counsellors, psychologists, psychiatrists, internists), leaving the question of whether psychosocial interventions performed by GPs might be effective...There are several reasons why GPs could use the knowledge and skills to perform psychosocial interventions in primary care. As patients come to their GP first with their health concerns, it is desirable that all potential treatment options in primary care are considered before a patient is referred to specialist care...Secondly, many GPs already take the time to support their distressed patients without having adequate tools to structure these extended visits...Thirdly, the degree of success of any kind of psychosocial intervention depends largely on the trust one places in the care-provider...Finally, it is recommendable to endorse a holistic approach of patient care (Richardson 1989)." *Objective* "we aimed to present a systematic review of all the available literature addressing the effectiveness of psychosocial interventions by general practitioners."

Rice VH, Stead LF. Nursing interventions for smoking cessation[225]:
Background "The aim of this review is to examine and summarize randomized clinical trials where nursing provided smoking cessation interventions. The review therefore focuses on the nurse as the intervention provider, rather than on a particular type of intervention." *Discussion* "The US Public Health Service clinical practice guideline 'Treating Tobacco Use and Dependence' (AHRQ 2000) used logistic regression to estimate efficacy for interventions delivered by different types of providers. Their analysis did not distinguish among the non-physician medical healthcare providers, so that dentists, health counsellors, and pharmacists were included with nurses. The guideline concluded that these providers were effective (Table 15, OR 1.7, 95% CI 1.3 to 2.1)." *Conclusion* "The evidence suggests that brief interventions from nurses who combine smoking cessation work with other duties are less effective than longer interventions with multiple contacts, delivered by nurses with a role in health promotion or cardiac rehabilitation."

Box 3.16 Continued…

Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol[224]:

Background "Dietitians are specifically trained and motivated to provide high quality dietary advice. Dietitians have a variety of different approaches available in order to provide advice and information that is appropriate for an individual patient. Due to the limited number of dietitians and the large proportion of the population who are at risk from, or have, coronary heart disease, much of the dietary advice is given by physicians and nurses rather than by dietitians with extensive nutrition training (Summerbell 1996). The effectiveness of dietary advice given by dietitians compared with other health professionals or self-help resources is unknown. Knowledge of the relative effectiveness would inform policy decisions on the best way to manage raised blood cholesterol in the general population." *Objective* "Primary objective: The review aimed to answer the following question: In adults, what is the relative efficacy of dietary advice given by a dietitian compared with another health professional, or using self-help resources in reducing blood cholesterol?"

EXAMPLES FROM OTHER REVIEWS

Doggett C, Burrett S, Osborn DA. Home visits during pregnancy and after birth for women with an alcohol or drug problem[259]:

Objective "Secondary objectives...were to examine the evidence for...person/s doing the visit:...counselor, nurse, or trained lay worker..." *Methods* "As home visiting encompasses a heterogeneous group of interventions...Subgroup analyses were therefore performed according to:...person/s doing the visit:...social worker, counselor, nurse, or trained lay worker..."

Hodnett ED, Fredericks S. Support during pregnancy for women at increased risk of low birthweight babies[249]:

Background "Debates have arisen regarding the relative benefits of 'professional' versus 'peer' support. Social support from a woman in one's community, who has a similar socioeconomic background and is experiencing similar life stresses, may be qualitatively different from support from a healthcare professional, who has broad professional knowledge and experience, but may not share the same socioeconomic background or life concerns, and who often provides other professional services as well as support. This Review includes studies of support by providers with varying backgrounds and qualifications." *Objectives* "Secondary objectives were to determine whether effectiveness of support was mediated by...type of provider (a healthcare professional or a lay woman)." *Methods* "A subgroup analysis was planned to compare support provided by lay women versus support by healthcare professionals, because another Review of support for childbearing women (Hodnett 2003) found differences in the effects of support by hospital staff (nurses, midwives) versus support by lay women." *Results* "Because there was only one trial in which the support was provided by lay women (Spencer 1989), and in another trial the support was provided by a multidisciplinary team that included lay women (McLaughlin 1992), the planned subgroup analysis was not performed. However, the results of these two trials were remarkably consistent with those of the other trials."

Ostelo RWJG, van-Tulder MW, Vlaeyen JWS, Linton SJ, Morley SJ, Assendelft WJJ. Behavioural treatment for chronic low-back pain[195]:

Discussion "In most of the studies included in this review, the qualifications of the care providers were not explicitly described. Therefore, we were not able to perform our pre-planned subgroup analysis of studies with qualified versus unqualified psychologists. We would suggest that in future trials the authors explicitly describe the qualifications and experience of therapists."

Ray KL, Hodnett ED. Caregiver support for postpartum depression[229]:

Summary "not enough evidence to show...any differences that may exist in the support provided by health professionals or lay people."

Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation[207]:

Discussion "There is still limited evidence from which to identify those elements of group therapy which are most important for success. In the main analyses there are too few studies to compare subgroups of studies according to content, provider or length."

Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents[181]:

Methods "Factors that may affect the results from individual studies were investigated using sensitivity analyses. This review proposed to investigate...differences between the person administering the intervention (e.g., nurse versus parent versus doctor)"

Box 3.17 Comments on the Use of Published Quality Assessment Tools

Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders[221]:
Methods "CCDAN Quality Rating Scale (Moncrieff 2001) criteria were used to determine external validity and study quality. This scale had 23 items with a maximum possible score of 46. Parameters included clarity of objectives, sample size, duration, power calculation, method of allocation, concealment of allocation, treatment description, blinding, source of subjects, use of diagnostic criteria, record of exclusions, sample description, blinding of assessors, assessment of compliance, side effects, withdrawals, description of outcome measures, adjustments for differences, inclusion of withdrawals in analysis, presentation of results, statistical analysis, justification of conclusions and declaration of interests. Each study was rated on 23 items to give a score ranging from 0 to 46." *Methodological Quality* "Some of the elements of the CCDAN scale were not relevant to this type of treatment research. There was no blinding of psychotherapy subjects and specific "side effects" were reported." *Discussion* "The studies were of variable quality...Manuals and adherence measures were not employed in each study calling into question the quality of psychotherapy provided. Therapist experience was in question in many studies, raising the chance that the therapy was not provided in an optimal fashion...The CCDAN Quality Rating System we used did not include ratings on these parameters, which were relevant to the interpretation of psychotherapy study quality"

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Methodological Quality "None of the studies was double blind. It is rarely possible to blind participants or therapists to psychological interventions, so the application of trial quality assessment tools, such as the Oxford scale (Jadad 1996) nor other widely used quality scales (Juni 2001), was not appropriate."

Jeffery DP, Ley A, McLaren S, Siegfried N. Psychosocial treatment programmes for people with both severe mental illness and substance misuse[190]:
Methodological Quality "The Jadad scoring focuses on not only randomisation but also blindness of rating at outcome and description of the reasons for leaving the study early...In retrospect it may have been inadvisable to use this scoring system for these sorts of studies where blinding is rare and particularly difficult. Nevertheless this scale was validated using trials from several specialities including mental health (Moher 1998)."

Ostelo RWJG, van-Tulder MW, Vlaeyen JWS, Linton SJ, Morley SJ, Assendelft WJJ. Behavioural treatment for chronic low-back pain[195]:
Methods "The methodological quality of the RCTs was independently assessed...using a criteria list (Table 01) recommended by the Cochrane Back Review Group (van Tulder 1997b; van Tulder 2003)...As it is difficult to blind patients for behavioural treatment, we redefined the criterion regarding the blinding of patients. If blinding was not feasible, item 4 of the criteria list was scored positive if the credibility of the treatments was evaluated and treatments were equally credible and acceptable to patients (Turk 1993)."

Rose S, Bisson J, Churchill R, Wessely S. Psychological debriefing for preventing post traumatic stress disorder (PTSD)[218]:
Methods "The third was a scale derived from Kenardy 1996a giving proposed quality standards for trials of psychological debriefing." *Methodological Quality* "The differences between the more general Moncrieff 2001 and the specific Kenardy 1996a scales reflect that fact that the Moncrieff 2001 scale emphasises general methodological issues relevant to all clinical trials, with a particular emphasis towards pharmacological trials, albeit relevant to psychiatry. The Kenardy 1996a scale gives more weight to specific issues concerning debriefing, and in particular the content of debriefing."

Box 3.18 Inclusion of Therapist Characteristics in Review Eligibility Criteria

GOOD EXAMPLES

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Eligibility "Counselling may be offered by a variety of professionals (e.g. counsellors, community psychiatric nurses (CPNs), practice nurses, social workers, clinical psychologists, GPs and health visitors). In this review, there were no specific inclusion or exclusion criteria related to professional background. However, formal counselling training was considered essential, to standardise expertise and practice. Only practitioners with a formal counselling qualification equivalent to BACP accreditation levels (www.bacp.co.uk/accreditation/index.html) were included in the review."

Littell JH, Popa M, Forsythe B. Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17[213]:
Background "As described by its developers (Henggeler 1998)...Treatment teams consist of professional therapists and crisis caseworkers, who are supervised by clinical psychologists or psychiatrists. Therapists are mental health professionals with masters or doctoral degrees; they have small caseloads and are available to program participants 24 hours a day, seven days a week." *Eligibility* "To be included in this review MST programs had to be licensed; other multisystemic treatments were not included."

Box 3.18 Continued...

OTHER EXAMPLES

Anderson CS, Hackett ML, House AO. Interventions for preventing depression after stroke[236]:
Eligibility "(2) A comparison between a psychological therapy and standard care (or attention control) for the prevention of depression associated with stroke...All interventions had to...be delivered by somebody with some explicitly stated training and supervision in therapies"

Baldwin C, Parsons T, Logan S. Dietary advice for illness-related malnutrition in adults[202]:
Eligibility "Dietary advice was defined as instruction in modification of food intake given with the aim of improving nutritional intake by a dietitian or other health care professional."

Barbato A, D'Avanzo B. Marital therapy for depression[211]:
Eligibility "Types of intervention: Inclusion criteria: 1. Marital therapy, for the purpose of this review, was defined as a structured psychological intervention in which a trained therapist met both partners in a couple, in regular sessions, with the explicit aim of modifying dysfunctional patterns of interaction."

Buckley LA, Pettit T. Supportive therapy for schizophrenia[222]:
Summary "As we were unable to find a widely accepted definition of supportive therapy, we developed our own...This includes interventions that require a trained therapist, such as supportive psychotherapy, as well as other interventions that require no training, such as 'befriending'."

Carr AB, Ebbert JO. Interventions for tobacco cessation in the dental setting[227]:
Eligibility "We included any intervention...which included a component delivered by a dentist, dental hygienist, dental assistant or office staff in the dental practice setting and any combination of these, as well as the same individuals providing intervention as part of a community effort...Interventions aimed at the training of dental health professionals were included."

den-Boer PCAM, Wiersma D, Russo S, van-den-Bosch RJ. Paraprofessionals for anxiety and depressive disorders[226]:
Eligibility "Inclusion criteria: Randomised controlled trials that used symptom measures, and compared the effects of any kind of psychological treatment given by paraprofessionals with psychological treatments given by professionals, or with waiting list or placebo condition."

Dennis CL, Creedy D. Psychosocial and psychological interventions for preventing postpartum depression[188]:
Eligibility "Any form of standard or usual care compared to a variety of non-pharmaceutical interventions...by a professional (nurse, midwife, childbirth educator, physician) or lay person (a specially trained woman from the community, a student)".

Doggett C, Burrett S, Osborn DA. Home visits during pregnancy and after birth for women with an alcohol or drug problem[259]:
Eligibility "Home visits...by teams or individuals consisting of doctors (obstetricians, general practitioners or pediatricians), nurses (midwives, drug and alcohol workers or early childhood nurses), social workers, counselors, or trained lay people...Studies that detailed timing of visits, frequency of visits, the type of home visitors, the interventions and co-interventions were included."

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Eligibility "No restrictions were placed on where or who delivered the therapy."

Hackett ML, Anderson CS, House AO. Interventions for treating depression after stroke[239]:
Eligibility "(3) a comparison between a psychological therapy and standard care for the treatment of depression associated with stroke…All interventions had to...be delivered by somebody with some explicitly stated training and supervision in therapies."

Hay PJ, Bacaltchuk J, Stefano S. Psychotherapy for bulimia nervosa and binging[191]:
Eligibility "Cognitive behaviour psychotherapy: For the purpose of this review, this is a psychotherapy that uses the specific techniques and model, but not necessarily the number of sessions or specialist expertise, of the cognitive and behavioural therapy therapy for bulimia nervosa as described by Fairburn and colleagues (CBT-BN; Fairburn 1993b)...In the analyses comparing CBT to pure self-help, guided self-help when guided by someone with some expertise, is thus "allowed" as CBT"

Hodnett ED, Fredericks S. Support during pregnancy for women at increased risk of low birthweight babies[249]:
Background "The programs may be delivered by multidisciplinary teams of health professionals, by specially trained lay workers, or by a combination of lay and professional workers. This Review includes all acceptably controlled trials of such programs."

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck Cv. Psychosocial interventions delivered by general practitioners[228]:
Objectives "3) To assess the effectiveness of psychosocial interventions by general practitioners compared to the reference treatment (whether 'usual care' or another experimental intervention) by reviewing the clinical outcomes of the selected studies." *Description of Studies* "In a second included study (Lidbeck 1997), the

Box 3.18 Continued…

therapist performing all interventions was a physician trained in family medicine as well as internal and social medicine who worked in a preventive medicine unit in primary care, which raised our doubts whether this therapist could be classified as a typical general practitioner. However, since the study formally met our inclusion criteria, we decided to include the study in the review."

James A, Soler A, Weatherall R. Cognitive behavioural therapy for anxiety disorders in children and adolescents[199]:
Eligibility "Types of intervention: Manualised CBT of at least eight sessions provided by trained therapists under regular supervision."

Joy CB, Adams CE, Rice K. Crisis intervention for people with severe mental illnesses[201]:
Eligibility "i. Crisis intervention: any type of crisis-orientated treatment of an acute psychiatric episode by staff with a specific remit to deal with such situations, in and beyond 'office hours'."

Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation[210]:
Background "In this review, we therefore specifically exclude counselling provided by doctors or nurses during the routine clinical care of the patient, and focus on smoking cessation counselling delivered by specialist counsellors." *Eligibility* "This review specifically excludes studies of counselling delivered by doctors and nurses as part of clinical care, which are covered in separate reviews (Rice 2004; Silagy 2004)."

Martinez DP, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus[200]:
Eligibility "Cognitive behavioural therapy (...by a qualified practitioner) versus no treatment or other treatments."

Perkins SJ, Murphy R, Schmidt U, Williams C. Self-help and guided self-help for eating disorders[248]:
Eligibility "(b) Guided self-help: For the purpose of this review, this refers to the above self-help definition, plus contact with a 'therapist' who may be a mental health professional or lay person."

Ray KL, Hodnett ED. Caregiver support for postpartum depression[229]:
Eligibility "All types of professional and/or social support including emotional support, counselling, tangible assistance and information..."

Rees K, Bennett P, West R, Davey SG, Ebrahim S. Psychological interventions for coronary heart disease[176]:
Eligibility "All non-pharmacological psychological interventions delivered by health care workers with specific training in these techniques were considered."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Eligibility "Interventions could have been delivered by psychologists, counsellors, medical staff, nurses, occupational therapists or other health professionals..."

Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol[224]:
Eligibility "Accepted interventions included dietary advice given by a dietitian or a nutritionist compared with another health professional (e.g. doctor or nurse) or self-help resources. Nutritionists as well as dietitians have been included as in different settings and different countries the terms dietitian and nutritionist may both be used to describe a health professional trained to give dietary advice."

Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents[181]:
Eligibility "Interventions administered by any qualified health-care professional (i.e., doctor, nurse, psychologist, technician), family member, caregiver, or by the child him/herself after being trained by a parent or professional, or both were included."

Woods B, Spector A, Jones C, Orrell M, Davies S. Reminiscence therapy for dementia[220]:
Eligibility "Types of intervention:...Only trials where...sessions were led by professional staff (psychologists, occupational therapists, nurses etc.) or by care-workers with training from professional staff were included."

Yorke J, Fleming SL, Shuldham CM. Psychological interventions for adults with asthma[182]:
Eligibility "Any type of psychological intervention used in the treatment of asthma in adults was considered for this review...These interventions will be delivered by a trained practitioner or in consultation or supervision by a trained practitioner."

Box 3.19 Examples of Reporting of Therapist Samples in the Review Text

Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders[221]:
Description of Studies "Eleven of these studies described using experienced therapists, but it was often unclear whether the therapists were experienced in the specific brief therapy approach versus other psychotherapy models."

Anderson CS, Hackett ML, House AO. Interventions for preventing depression after stroke[236]:
Description of Studies "The interventions were delivered by a variety of trained professionals, including specialist nurses (Forster 1996; House 2000) and a mixed team of therapists (Goldberg 1997)."

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Results "Types of practitioner: A range of practitioners offered a range of counselling interventions. In seven of the trials, all the professionals had the necessary qualifications and experience to be accredited by the BACP (Boot 1994, Harvey 1998, Hemmings 1997, Friedli 1997, King 2000, Simpson 2000, Barrowclough 2001). In one trial, it was not clear whether all the included counsellors met the criteria for BACP accreditation (Chilvers 2001), although correspondence with the authors indicated that a significant proportion did, and all were highly experienced."

Doggett C, Burrett S, Osborn DA. Home visits during pregnancy and after birth for women with an alcohol or drug problem[259]:
Results "Three studies (Black 1994; Butz 1998; Quinlivan 2000) used nurses to provide home visits...No study reported using trained social workers to provide home visits...Dakof 2003 provided a manualised home-based, goal-orientated program administered by trained ’black’ specialists with prior experience in drug treatment services...Two studies (Grant 1996; Schuler 2000) reported home visits by trained lay workers...No study reported using a multidisciplinary team to provide home visits."

Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents[183]:
Methodological Quality "Therapist training and competence: The trials employed a variety of therapists ranging from undergraduate assistants to experienced psychological and medical personnel, but mainly graduate trainees in clinical psychology. Other trials employed non-psychologists specifically trained for the trials (eg, school nurses and teachers) to deliver structured interventions. The level of therapist training was not stated in six trials. Only three trials explicitly mentioned that therapists received supervision during the trials. This, coupled with the general failure to note whether checks on adherence were made, must be considered a weakness when judging the overall quality of the trials."

Hunot V, Churchill R, Silva-de LM, Teixeira V. Psychological therapies for generalised anxiety disorder[185]:
Description of Studies "The therapists employed to conduct psychological therapy treatments were predominantly qualified professionals, consisting of clinical psychologists (n=11), doctoral/senior/advanced level CBT therapists (n=5) and experienced therapists/therapists (n=5). A small number of studies used graduates/advanced graduates (n=3). One study did not describe the therapists used to conduct the treatment (Lavallee 1993)."

James A, Soler A, Weatherall R. Cognitive behavioural therapy for anxiety disorders in children and adolescents[199]:
Methodological Quality "Therapists were mostly post doctorate psychologists"

Ostelo RWJG, van-Tulder MW, Vlaeyen JWS, Linton SJ, Morley SJ, Assendelft WJJ. Behavioural treatment for chronic low-back pain[195]:
Description of Studies "There were 15 RCTs that specifically mentioned the qualification of therapists and six RCTs that did not mention the therapists' qualifications (Altmaier 1992; Bru 1994; Donaldson 1994; Lindström 1992; Newton-John 1995; Stuckey 1986). An example of sufficient description of qualifications of therapists was 'psychologist who had had five years of experience with chronic pain patients since completing his clinical qualifications' (Nicholas 1991). An example of insufficient description was 'a physical therapist' (Lindström 1992)."

Rice VH, Stead LF. Nursing interventions for smoking cessation[225]:
Description of Studies "We determined whether the nurses delivering the intervention were providing it alongside clinical duties that were not smoking related, were working in health promotion roles, or were employed specifically as project nurses. Of the high intensity intervention studies, five used nurses for whom the intervention was a core component of their nursing role (Hollis 1993; DeBusk 1994; Allen 1996; Carlsson 1997; Terazawa 2001). In six studies the intervention was delivered by a nurse specifically employed by the project (Taylor 1990; Rice 1994; Rigotti 1994; Miller 1997; Lewis 1998; Canga 2000). In three of these, the same nurse provided all the interventions (Rigotti 1994; Lewis 1998; Canga 2000). In only three studies were intensive interventions intended to be delivered by nurses for whom it was not a core task (Lancaster 1999; Bolman 2002; Curry 2003). In the last of these the intervention was given either by paediatric nurses or by health educators. All the low intensity interventions were delivered by primary care or outpatient clinic nurses.”

Box 3.20 Discussion of Generalisation of Results to Therapist Populations

EXPLICIT COMMENTS

Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care[37]:
Discussion "The practices and GPs recruited to the studies were 'volunteers' rather than a random sample. The doctors who participated may have been particularly interested in the research question and may have used therapeutic techniques to a greater extent than is usual, thus reducing the additional effect of counselling (Friedli 1997). In addition, in one trial (Hemmings 1997), GPs participated in an Action Learning programme, in which they learned about counselling and counselling skills. This may have effected their consultation style and referral practices." *Conclusion* "The results can only be generalised to similar patients and counsellors. This means that the evidence is restricted to counsellors with BACP accreditation or equivalent."

Eustice S, Roe B, Paterson J. Prompted voiding for the management of urinary incontinence in adults[217]:
Discussion "External validity has been weakened due to the use of research staff to implement the intervention in four of the trials (Hu 1989; Ouslander 2005; Schnelle 1989; Schnelle 2003). The other trials have demonstrated that it is possible to use direct caregivers during the trial period (Engberg 2002; Linn 1995; Smith 1992; Surdy 1992; Schnelle 1983). Reliability checks were performed for the wet checks in seven trials, which may be partly responsible for the compliance of staff with the programme. Two trials did not report reliability checks (Linn 1995; Smith 1992). Without robust trials that address these issues, our understanding of the factors that influence the successful management of urinary incontinence will remain unclear. Nevertheless, these trials are important for exposing the multidimensional aspects of managing incontinence in a frail, elderly population. Therefore, this work adds to the literature on behavioural treatment of urinary incontinence, but the body of knowledge remains incomplete, especially within the nursing home environment."

Furukawa TA, Watanabe N, Churchill R. Combined psychotherapy plus antidepressants for panic disorder with or without agoraphobia[172]:
Discussion "generalisability of the present findings beyond specialist psychiatric settings is not straightforward. Only one study (Sharp 1996) was conducted in the primary care setting but in this study patients were seen in their local GP clinics but by qualified clinical psychologists. Only two studies in this review assessed a psychological approach other than CBT. Where CBT therapists were not available, the research evidence, as accumulated and presented in this systematic review, would not be readily applicable to clinical practices."

Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck CP van. Psychosocial interventions delivered by general practitioners[228]:
Discussion "These findings should be interpreted with considerable caution: the two studies on PST were conducted by the same research team and groups consisting of only 30 to 40 patients were treated by a small number of experienced research GPs…This finding exposes an interesting problem in primary care and general practice research: in order to have a large sample size, investigators are often forced to recruit a large number of GPs who can select patients for the trial and perform the intervention. However, there are many objections to having a large number of GPs actively participating in the trial: patient recruitment by GPs is usually slow, accurate training and supervision of a large number of GPs is time-consuming and costly and GPs do not build experience if they only apply the intervention to a small number of patients (Van der Windt 2000). A solution to this problem is the deployment of a small number of research GPs who have (unfamiliar) patients assigned to them for the purpose of the trial. These research GPs tend to be highly motivated, highly trained and highly experienced. Consequently, findings from studies in which a small sample of research GPs performs the intervention should be interpreted with caution, since research GPs might not represent the typical GP and treatment effects might therefore be an overestimation of the effects in daily practice. In this review, of the six studies that have a small number of GPs delivering the intervention, four studies use research GPs instead of the regular GP of patients. Although treatment by the regular GP versus treatment by an unfamiliar research GP was not associated with outcome, these circumstances make the available evidence even harder to interpret. Of course, this highly qualitative analysis of comparing study characteristics in relation to outcome lacks a clear validity basis due to the small number of selected studies. Nevertheless, the potential influence of factors like these, especially the influence GP-factors (number, type, training, experience), should be explored in future research." *Conclusion* "Problem-solving treatment by highly experienced GPs seems a promising tool in the treatment of depressed patients, although the effectiveness of this intervention by regular GPs in routine care remains to be demonstrated."

Hunot V, Churchill R, Silva-de LM, Teixeira V. Psychological therapies for generalised anxiety disorder[185]:
Discussion "Since 79% of therapists and counsellors in UK primary care are person-centred or integrative in theoretical orientation and CBT is only practiced by 10% of therapists (Stiles 2006), the evidence produced in this review could be regarded as of limited applicability. Furthermore, in the majority of studies, the therapists employed were highly qualified and experienced practitioners, who may not be representative of practitioners employed in real world clinical settings."

Box 3.20 Continued…

Jones C, Cormac I, Silveira-da-Mota-Neto-JI, Campbell C. Cognitive behaviour therapy for schizophrenia[197]:
Conclusion "For clinicians: Presently, cognitive behavioural therapy is a scarce commodity, often provided by highly skilled and experienced therapists. Therefore, its application in day-to-day practice may be restricted by the availability of suitably qualified practitioners. The present data provides little indication of how effective cognitive behavioural therapy procedures might be when they are applied by less experienced practitioners." *Summary* "Cognitive behavioural therapy (CBT) is one of the talking therapies that is suggested to be of value to people with schizophrenia. This review suggests that it may well be of value, at least in the short term. Cognitive behavioural therapy should be further evaluated in various clinical settings and comparing effects for both expert and less skilled practitioners."

Silagy C, Lancaster T, Stead L, Mant D, Fowler G. Nicotine replacement therapy for smoking cessation[261]:
Discussion "Nicotine gum and transdermal patches were more effective when offered to volunteer smokers recruited from the community or those attending specialized clinics than if offered to smokers in primary care. These findings are likely to be partly explained by the high motivation to quit among many of the smokers in the community who volunteer for trials in response to media advertisements and, similarly, among those participants who are recruited as a result of their attendance at specialized smoking cessation clinics. The latter group also have access to trained therapists who specialize in assisting smokers to quit. However, given the limited number of specialized smoking cessation clinics, access will be restricted to a small proportion of smokers wanting help to quit. In contrast, most of the smokers recruited into trials conducted in primary care settings were unselected, and hence may be less motivated to quit. In addition, the treating physician or practice nurse had frequently received little training in smoking cessation skills. As a result, compliance with NRT among smokers treated in primary care is reported to be lower than in other settings (Lam 1987)."

IMPLICIT COMMENTS

Buckley LA, Pettit T. Supportive therapy for schizophrenia[222]:
Conclusion "Future trials should clearly explain whether practitioners who deliver supportive therapy have been specifically trained, and if so how. It may make the results more applicable if the therapists are trained but in the context of routine career development, rather than specific highly-trained specialised practitioners."

den-Boer PCAM, Wiersma D, Russo S, van-den-Bosch RJ. Paraprofessionals for anxiety and depressive disorders[226]:
Conclusion "Significant questions remain about the conditions under which paraprofessionals can be effective. Most studies mention some selection, training and supervision of paraprofessionals. If paraprofessionals, volunteers or patients, can be effective therapists (with no training or minor initial training), or can offer support because of their personal experience with the underlying problem, this will bring psychological treatment within the scope of psycho-education or education alone. The evidence presented so far may justify the development of new programs incorporating paraprofessionals."

Gold C, Heldal TO, Dahle T, Wigram T. Music therapy for schizophrenia or schizophrenia-like illnesses[215]:
Discussion "The specific techniques of music therapy, including, among others, musical improvisation and the discussion of personal issues related to the musical processes, require specialised music therapy training. Both training courses and qualified music therapists are available in many countries, but in some countries there may be a need for development of good quality training."

Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder[214]:
Discussion "When applying the results of this review to practice, it is important to note that the application of music therapy requires an academic and clinical training in music therapy. Trained music therapists are available in many countries. Training courses in music therapy teach not only the clinical music therapy techniques as described in the background of this review, but also aim at developing the therapist's personality and clinical sensitivity, which is necessary to apply music therapy responsibly. Academic training courses in music therapy exist in many countries, and information is usually available through the professional associations." *Abstract* "When applying the results of this review to practice, it is important to note that the application of music therapy requires specialised academic and clinical training."

Joy CB, Adams CE, Rice K. Crisis intervention for people with severe mental illnesses[201]:
Discussion "It is unfortunate that no data are available for staff satisfaction. Issues such as staff recruitment, despondency and burnout are essential to the successful implementation of home care packages. Several of the studies mentioned these as notable problems affecting the running of the project. If such problems were prominent in these usually well-resourced and well-motivated research teams, they may amount to insurmountable obstacles to the implementation of similar projects in routine psychiatric settings."

Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation[210]:
Conclusion "Implications for practice: Counselling interventions given outside routine clinical care, by smoking cessation counsellors including health educators and psychologists, assist smokers to quit."

Box 3.20 Continued…

Morriss RK, Faizal MA, Jones AP, Williamson PR, Bolton C, McCarthy JP. Interventions for helping people recognise early signs of recurrence in bipolar disorder[173]:
Discussion "Perry 1999 was the only trial that used therapists with little previous experience to deliver the intervention. Therefore, successful EWS interventions seem to require around 12 sessions of therapist time and involve therapists of high competency...The one EWS intervention that used a less experienced and, therefore, less expensive therapist only showed a benefit against manic type recurrences and function, without any effect against depressive type recurrence. The relative cost-effectiveness of interventions involving more experienced versus less experienced therapists is a topic for further research." *Conclusion* "The means by which EWS interventions could be efficiently delivered within existing health service systems is not clearly established."

O'Connor AM, Stacey D, Entwistle V, Llewellyn TH, Rovner D, Holmes RM, Tait V, Tetroe J, Fiset V, Barry M, Jones J. Decision aids for people facing health treatment or screening decisions[254]:
Conclusion "However, several conditions are necessary to implement decision aids in practice. These include:...b) practitioners willing to try decision aids in their practice;...d) practitioners and health care consumers who are skilled in shared decision making."

O'Kearney RT, Anstey KJ, von SC. Behavioural and cognitive behavioural therapy for obsessive compulsive disorder in children and adolescents[194]:
Discussion "Applicability of results: The BT/CBT interventions reviewed are similar, with three of the four using a standard protocol for the BT/CBT treatment of OCD in children and adolescents. While the protocol and manuals are readily available (March 1998), training and supervision in its delivery are less accessible." *Conclusion* "An equally important area of research arising from the results would be an examination of how well BT/CBT could be disseminated and implemented in non-specialist centres or non-academic settings, as often the main limitation to offering BT/CBT is availability of skilled therapists."

Perkins SJ, Murphy R, Schmidt U, Williams C. Self-help and guided self-help for eating disorders[248]:
Conclusion "It remains uncertain whether and how much guidance is needed and from whom."

Ray KL, Hodnett ED. Caregiver support for postpartum depression[229]:
Conclusion "As with other studies of support for childbearing women, e.g. Cochrane Review 'Caregiver support for women during childbirth' (Hodnett 2001), questions remain about the relative benefits of social (for example, lay person) versus professional (health visitor, nurse, midwife) support."

Rice VH, Stead LF. Nursing interventions for smoking cessation[225]:
Conclusion "Additionally, controlled studies are needed that carefully examine the effects of 'brief advice by nursing' as this type of professional counselling may more accurately reflect the current standard of care."

Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis[180]:
Conclusion "There are a number of issues that need to be addressed in relation to the delivery of psychological treatment. For example,…What is the potential for professions other than psychologists, such as nurses and occupational therapists, to conduct psychology-based interventions?...Two of the studies qualifying for this review used interventions that were delivered by nurses. Psychologists working in the acute hospital setting in the NHS are in short supply. Training and supporting other health professionals to deliver psychology-based interventions, or to deliver interventions that incorporate psychological principles, could be useful. This would need to be corroborated by research evidence."

Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol[224]:
Conclusion "Further work is needed on which elements of dietary advice make it effective, e.g….level of belief of practitioner, level of training of practitioner..."

4 ILLUSTRATIVE EXAMPLE: COUNSELLING IN PRIMARY CARE

4.1 Introduction

The Cochrane reviews included in Chapter 3 provided the sampling frame for selecting an illustrative example for the remainder of the thesis. Bower and Rowland's[37] review, on the clinical and cost-effectiveness of counselling in primary care, was chosen on the basis of the following three criteria. Firstly, the review topic had to be clearly oriented towards psychotherapy. Secondly, the data structure of the studies involved in at least one meta-analysis had to highlight the heteroscedasticity issues in a relatively intuitive manner. Finally, the individual-patient-data (IPD) had to be accessible. The aim of this chapter was to introduce the example, providing an overview of the Cochrane review, the randomised trials included, the patient-level data obtained, and the practical issues arising from the information that was available.

4.2 The Cochrane Review

4.2.1 Background

The principal point of contact for patients presenting in primary care is their general practitioner (GP) and associated primary care team. One in three are estimated to be affected by mental health problems[290]. By 2005, there were more people receiving benefits for incapacity through mental ill-health than the total number unemployed[290]. The case for providing psychological therapies, including counselling, within the NHS has been made recently by five leading mental health charities[291], as part of a wider drive to improve access to these forms of treatment[290, 292-294]. Reimbursement for the cost of employing counsellors in general practice has been available since 1990[295]. The intervening period has seen a rapid rise in counselling in primary care[296, 297], with half the general practices in England estimated to have a counsellor attached by 2000[298]. Counselling is defined in this context as "a systematic process which gives individuals an opportunity to explore, discover and clarify ways of living more resourcefully, with a greater sense of well-being. Counselling may be concerned with addressing and resolving specific problems, making decisions, coping with crises, working through conflict, or improving relationships with others." (p.9)[299]

Table 4.1 What is essential and desirable in a counsellor in primary care

Criteria	Essential	Desirable
Education and professional qualifications	450 hours training	BACP accreditation or equivalent
Knowledge	One theoretical approach to counselling	Variety of counselling theories and methods
	Psychosomatic disease and psychology of chronic or terminal illness	Psychotrophic drugs and their side effects
	BACP code of ethics – particularly about confidentiality	Psychopathology by visiting admission unit of psychiatric hospital
Experience	250 hours supervised counselling over 2 years	At least 300 hours gained over at least 3 years
Personality	Dependable	Aware of boundaries around punctuality
	Considered approachable by a wide range of patients?	Friendly
Physical attributes	Good enough health and sufficient sight and hearing not to make special demands on clients	Able to work under pressure and to monitor and manage own stress level
Special circumstances	A constructive member of a multidisciplinary team	Understanding of culture of medical settings and willingness to develop appropriate counselling skills among team members

Reproduced from Table 1.3 in Bond T. The nature and role of counselling in primary care. In: Keithley J., Bond T, Marsh G (eds) *Counselling in Primary Care.* Oxford University Press: Oxford, 2002. Updated from The Counselling in Primary Care Trust criteria[300].

The background of counsellors working in this setting is variable[296]. While the NHS has not set specific training standards[301], the most widely accepted are those required for accreditation by the British Association for Counselling and Psychotherapy (BACP), leading to registration with the United Kingdom Register of Counsellors[296]. Professional recognition is acquired through membership of the Faculty of Healthcare Counsellors and Psychotherapists, the Association of Counsellors and Psychotherapists in Primary Care, or the Counselling Psychology Section of the British Psychological Society[296]. The essential and desirable criteria set out by the Counselling in Primary Care Trust were updated by Bond[296] and are given in Table 4.1. In contrast to the BACP requirement of 450 hours of supervised practice, these only regard 250 hours as essential. There is also no mention of having received personal counselling, which is necessary for BACP accreditation. In a move to professionalise counselling in primary care, guidelines for employment of counsellors[302], appropriate referral of patients[303], and good practice[304] have been disseminated. Concerns about the competency of practice counsellors[305-309] prompted Mellor-Clark *et al*[298] to conduct a national survey. Of the 1031 responding, 75% held the recommended counselling diploma and 32% also had relevant Masters or Bachelors degrees. Eighty-three percent had at least two years post-qualification experience, with 76% also having a minimum of two years in a medical setting. All but two counsellors stated having regular individual or group supervision.

Counselling is typically brief, usually involving 6 to 10 sessions, each of 50 minutes[310]. Additional sessions may be offered as necessary, providing the resource is available[296]. The counselling process is characterised by three stages, operating by means of the relationship between the counsellor and the patient[296]. The focus is initially on building trust, with the counsellor employing many strategies and techniques to this end. The counsellor encourages the patient to describe the situation that is affecting them and makes a systematic assessment. The emphasis then turns to creating changes which give the patient additional resources they can subsequently draw upon. The way this is done by the counsellor depends on the theoretical model they are applying. Finally, alternative means of using the resources are considered, put into action and reflected upon. Regular meetings of the primary care team ensure the counsellor is not working in isolation[296]. The curriculum for training doctors now includes training in counselling and communication skills[311]. As a consequence, the separation between the respective roles of the counsellor and the GP may be blurred to varying degrees.

The nature of GP referrals reflects the diversity of patients encountered in a practice, the counsellor's specific competencies[298], patient and GP choice, and the availability of other services[37]. It is usual for counsellors to apply eclectic therapeutic approaches for a wide range of social and clinical problems. These include depression, anxiety, and bereavement, relationship difficulties, stress, and adjustment to physical ill-health[296, 298, 301]. The popularity of counselling amongst GPs and patients justifies the need for an evidence-base supporting its use. The rise in its availability, together with current moves to increase access still further, makes this all the more important. In line with proposals to identify empirically supported therapies[312], psychotherapy trials tend to focus interest on discrete therapeutic approaches for distinct diagnostic problems[298]. This runs counter to clinical practice in this example, however. As such, any attempts to restrict the severity or range of patient referrals beyond that accepted within good practice guidelines limits the generalisability of research findings. Restrictions based on the theoretical models counsellors are able to apply, their professional background, or the nature of the therapeutic relationship, the number and frequency of sessions also would have implications for generalisation.

4.2.2 Review Methodology

Bower and Rowland[37] searched MEDLINE, EMBASE, PsycINFO and CINAHL, together with the general and specific Cochrane trials registers, using an explicit and detailed search strategy. Studies were considered eligible if they met the following inclusion criteria

i)	Source	Published prior to July 2005
ii)	Study design	Randomised controlled trial
iii)	Setting	Primary care, including at home if referral from GP
iv)	Patient characteristics	Psychological or psychosocial problems regarded as suitable for counselling, including situational or life-adjustment problems not leading to a formal diagnosis
v)	Counselling	Distinct and separate treatment, potentially based on a variety of theoretical models, given as a series of sessions following an assessment to generate a plan, but not including specialist counselling interventions
vi)	Control	Usual care representing a mixture of interventions patients would typically receive, including GP referral to NHS psychotherapy services
vii)	Care providers	Counselling provided by professionals, possibly from a variety of backgrounds, with formal training equivalent to the requirements for BACP accreditation
viii)	Outcomes	Self-report or interviewer-rated a. Mental health symptoms b. Social and occupational functioning c. Patient satisfaction d. Costs

and eight studies were included[313-320]. A variety of control interventions were used with counselling compared to usual GP care in six studies[314, 316-320], to cognitive behavioural therapy in two[313, 319] and to GP-prescribed generic antidepressant treatment in one[315]. Included studies were rated using the Moncrieff *et al*[278] scale for their methodological quality. Items relating to side effects and the blinding of subjects and assessors were excluded as they were not regarded to be relevant.

4.2.3 Original Analysis

The reviewers summarised clinical effectiveness quantitatively and cost effectiveness in the text. All the published meta-analyses made use of continuous outcome scales. As different measurement scales were used across studies, the summary statistics extracted from each study were all standardised mean differences (SMDs). Outcomes were collected over an extended follow-up period, categorised as short-term (1 to 6 months), long-term (7 to 12 months) or very long-term (>12 months). The principal meta-analysis compared counselling to usual care, using the short-term outcomes that measured the extent of mental health symptoms. This was supplemented by an analysis comparing counselling to all forms of GP care, and by others based on long-term and other outcomes. Additional analyses were performed assessing how robust the principal analysis was to the exclusion of studies with compromised concealment of random allocations and patients presenting with chronic symptoms. In each case, the primary analysis assumed a common underlying treatment effect across studies. Evidence of between-study heterogeneity in the treatment effect was investigated with a chi-square test and the I^2 statistic[321]. Sensitivity analyses were performed using the random-effects model, assuming that the population treatment effects were normally distributed across studies. None of the meta-analyses made any allowance for within-study clustering or between-arm heteroscedasticity.

4.2.4 Published Results

Bower and Rowland[37] highlighted a number of limitations to the quality of the studies. Problems with the allocation procedure were reported in two trials[314, 318] contributing to imbalance in the number of patients assigned to each arm. In both cases, difficulties with concealment were attributed to a single GP. Boot *et al*[314] chose to retain affected patients in analyses, while Hemmings[318] excluded them. Volunteer GP practices were used to recruit patients in all eight trials. None reported what proportion of eligible patients were randomised, and only four were regarded as having provided adequate information on the characteristics of patients[313, 315, 319, 320]. This placed restrictions on the assessment of external validity. Neither Barrowclough *et al*[313] nor Boot *et al*[314] reported a power calculation, and additional information had to be sought from Harvey *et al*[317] and Hemmings[318] in order to perform the meta-analyses. None of the Moncrieff *et al*[278] items were extended to encompass the precision, internal or external validity implications of therapist variation.

The principal meta-analysis gave a pooled estimate of -0.28 (95% CI -0.43 to -0.13) for the SMD in short-term outcome for counselling versus usual GP care. This indicates that counselling reduces mental health symptom scores by an average of around 0.3 standard deviations, when compared to usual GP care in the short term. Excluding the studies that reported compromised concealment of random allocations[314, 318] gave very similar results (pooled SMD -0.27, 95% CI -0.45 to -0.09). Excluding the study focused on patients with chronic symptoms increased the pooled SMD estimate to -0.36 (95% CI -0.53 to -0.19). The meta-analysis comparing counselling to all forms of GP care, including data from Chilvers *et al*[315], is given in Table 4.2. This analysis was chosen to illustrate the methods within the remainder of the thesis. It is broadly consistent with the principal meta-analysis.

Table 4.2 Counselling compared to all GP care, short-term mental health outcomes

	Counselling			No Counselling					
Trial	**N**	**Mean (SD)**		**N**	**Mean (SD)**		**w**	**SMD (95% CI fixed)**	
Boot 1994	67	6.21	(6.97)	41	10.56	(8.97)	12.5	-0.55	(-0.95 to -0.16)
Chilvers 2001	39	15.20	(11.60)	44	14.80	(10.05)	10.5	0.04	(-0.39 to 0.47)
Friedli 1997	59	11.70	(7.70)	51	15.60	(10.50)	13.6	-0.43	(-0.80 to -0.05)
Harvey 1998	77	7.29	(4.57)	38	8.23	(5.05)	12.9	-0.20	(-0.59 to 0.19)
Hemmings 1997	114	0.98	(0.66)	40	1.03	(0.82)	15.0	-0.07	(-0.43 to 0.29)
King 2000	62	11.50	(7.70)	62	17.20	(11.90)	15.1	-0.57	(-0.92 to -0.21)
Simpson 2000	82	16.00	(9.30)	79	16.00	(8.10)	20.4	0.00	(-0.31 to 0.31)
TOTAL (95% CI)	**500**			**355**			**100**	**-0.24**	**(-0.38 to -0.10)**

Test for heterogeneity: $\chi^2 = 11.29$ $df = 6$ $p = 0.08$ $I^2 = 46.8\%$

Test for overall effect: $z = 3.43$ $p = 0.0006$

Note: This meta-analysis was published in Bower and Rowland[37] as Analysis 02.01 (p.56); w = weight; SD = standard deviation; CI = confidence interval; SMD = standardised mean difference; the outcome for Chilvers 2001, Friedli 1997, King 2000 and Simpson 2000 was the Beck Depression Inventory (BDI), for Boot 1994 it was the General Health Questionnaire (GHQ), for Harvey 1998 it was the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D), and for Hemmings 1997 it was the Symptom Index (SI).

Heterogeneity of the sample variances across arms is evident for a number of trials, but most notably for King *et al*[319] and Friedli *et al*[316]. The sample variances were larger in the control arm in five[314, 316-319] of the seven trials. Unequal patient ratios favoured the counselling arm in three trials[314, 317, 318]. The trial sample sizes, while small, were all of a comparable magnitude, exceeding 30 per arm. While only marginally significant, there is some evidence of heterogeneity in the SMDs between trials. Moderate effect sizes were observed in King *et al*[319], Boot *et al*[314] and Friedli *et al*[316], but no discernible effects were detected in Simpson *et al*[320], Chilvers *et al*[315] and Hemmings[318]. The 95% confidence intervals included moderate to large effects in all but Chilvers *et al*[315].

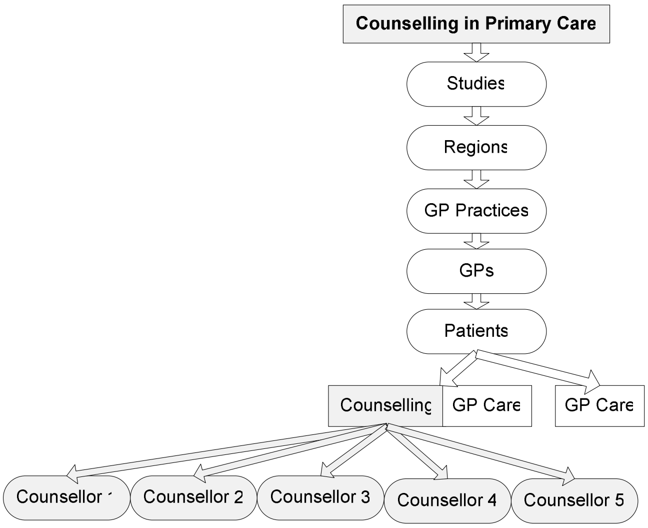
4.3 The Randomised Trials

4.3.1 Trial Methodology

The research question implicit in all seven trials[314-320] was whether counselling given by qualified counsellors in addition to GP care is more effective than GP care in reducing psychological distress. Interest was therefore in *packages* of therapeutic approaches and care provider characteristics. While six trials[314-319] framed the comparison in terms of counselling versus GP care, it is possibly more accurate to consider GP care a co-intervention, with the focus being on counselling versus no counselling. Patients were recruited through a GP in their local GP practice. The pre-existing allocation of GPs to patients suggests that the GP could be considered a patient characteristic, and hence a potential stratification factor. None of the trials stated that GPs were crossed with intervention arms, but this is a reasonable assumption in the circumstances. All seven trials individually randomised the treatment packages to patients. Patients expressing a strong preference could be assigned their preferred treatment in Chilvers *et al*[315] and King *et al*[319] under a patient preference design[322]. King *et al*[319] modified the allocation procedure mid-recruitment to add a third option of a two-way randomisation between counselling and cognitive behavioural therapy. The meta-analyses reported by Bower and Rowland[37] excluded these additional arms.

The region, GP practice, GP and counsellor were all potential sources of within-study clustering, giving rise to the idealised data structure depicted in Figure 4.1. Each trial was carried out in one or more regions within the United Kingdom. A number of GP practices were selected within each region, with one or more GPs nested within each practice. The patients in each trial therefore constitute clustered samples of patients within the United Kingdom. Once selected, patients were then randomly allocated an intervention; with those assigned counselling in turn allocated a counsellor. In clinical practice, counsellors would already be attached to one or more GP practices within a region. Those in Harvey *et al*[317] and Simpson *et al*[320] were already in post, but those in Boot *et al*[314] and Hemmings[318] were recruited and placed for the trial. The counsellors in Friedli *et al*[316] were peripatetic, while those in Chilvers *et al*[315] and King *et al*[319] came from a mixture of sources. As a consequence, some of the counsellors in some of the trials were crossed with GP practices, while others were nested. The cluster sampling of patients, together with the non-random allocation of counsellors to patients, could lead to clustering, regardless of whether or not patients attended treatment. Blinding providers of co-interventions to the intervention arm justifies the assumption that the underlying treatment-related clustering is homogeneous across arms. As none of the GPs were blinded, the possibility remains that treatment effects were heterogeneous across GPs due to performance bias. The between-arm differences in the nature of GP care in Chilvers *et al*[315] suggest that GP care is part of the treatment package, rather than being a co-intervention. This questions the relevance of performance bias for the GPs. For simplicity, however, the focus was placed on the counsellor-related clustering in subsequent chapters.

Figure 4.1 Idealised Data Structure for Counselling in Primary Care



The trial designs, patient populations, treatments and data collected are summarised in Table 4.3. In accordance with clinical practice, the patient populations sampled in Boot *et al*[314], Friedli *et al*[316], Harvey *et al*[317] and Hemmings[318] were broad, determined by GP referrals. In contrast, criteria were restricted in Chilvers *et al*[315], King *et al*[319] and Simpson *et al*[320] to patients with clinical depression, neither the severity or duration of which being reflective of referrals in clinical practice. The counselling provided varied across the trials, both in terms of the degree of standardisation and theoretical model applied. Between-study differences in the degree of treatment standardisation and the breath of the patient population may contribute to between-study heterogeneity in the size of the clustering effect because clustering is indexed by the ratio of the between-counsellor and the total variances.

The timing of the short-term outcome ranged from six weeks[314] to six months[320]. As

Table 4.3 Trial Characteristics

Trial	Design	Patients	Intervention	Control	Data Collection	
Boot *et al* 1994	1) Parallel-group RCT 2) Treatments randomly allocated to patients	1) Broad range 2) Acute but not severe psychological or psychosocial problems 3) GP referral	1) Generic counselling 2) No standardisation or manual 3) BAC accredited/accreditable counsellors 4) 6 1-hr weekly sessions	Usual GP care	1) GHQ-28 2) Patient satisfaction 3) Psychotropic drug use 4) GP Referral to outside agencies	1) Baseline 2) 6 wks
Chilvers *et al* 2001	1) Parallel-group RCT 2) Treatments randomised to patients using stratified block randomisation with GP practices as strata 3) Additional preference arms	1) Major depression 2) Research Diagnostic Criteria (Spitzer et al, 1978)	1) Generic counselling 2) No standardisation or manual 3) Counsellors with 2000+ hrs of supervised experience or in role already 4) 6 50-min weekly sessions	Routine GP antidepressant drug treatment (written protocol for guidance)	1) BDI 2) Global outcome 3) Time to remission 4) Research diagnostic criteria 5) SF-36 6) Costs	1) Baseline 2) 8 wks 3) 12 mths
Friedli *et al* 1997	1) Parallel-group RCT 2) Treatments randomised to patients using block randomisation	1) Broad range 2) Emotional difficulties GP regarded suitable	1) Non-directive psychotherapy (Rogerian model) 2) Standardised training & verification of therapy delivered 3) BAC accreditable counsellors 4) 6/12 50-min weekly sessions	Usual GP care (discouraged from referring to counsellors)	1) BDI 2) BSI 3) CIS-R 4) SAS-M 5) Patient satisfaction 6) Costs	1) Baseline 2) 3 mths 3) 9 mths
Harvey *et al* 1998	1) Parallel-group RCT 2) Treatments randomised to patients using block randomisation, ratio 2:1	1) Broad range 2) Emotional or relationship problems 3) GP referral	1) Generic counselling 2) No standardisation or manual 3) BAC accredited or diploma-level counsellors 4) 6 50-min weekly sessions	Usual GP care	1) HADS 2) COOP/WONCA 3) Delighted/terrible faces 4) SF-36 (Swansea only) 5) Costs	1) Baseline 2) 4 mths
Hemmings 1997	1) Parallel-group RCT 2) Treatments randomised to patients, ratio 2:1	1) Broad range 2) Appropriate referrals negotiated between GPs and counsellors	1) Generic Counselling 2) No standardisation or manual 3) BAC accredited counsellors	Usual GP care	1) Symptom Index 2) IIP-32 3) Repertory Grids 4) Patient satisfaction	1) Baseline 2) 4 mths 3) 8 mths
King *et al* 2000	1) Parallel-group RCT 2) Treatments randomised to patients using stratified block randomisation with severity (BDI score) as strata 3) Additional intervention (CBT), 2-way randomisation, and preference arms	1) Depression or depression/ anxiety 2) GP diagnosis/referral 3) >= 14 on BDI	1) Non-directive counselling (Rogerian model) 2) Manual & verification of therapy delivered 3) BAC accreditable counsellors 4) 6/12 sessions	Usual GP care (discouraged from referring to counsellors or prescribing antidepressants)	1) BDI 2) BSI 3) SAS-M 4) Patient satisfaction 5) Costs	1) Baseline 2) 4 mths 3) 12 mths
Simpson *et al* 2000	1) Parallel-group RCT 2) Treatments randomised to patients using random number tables	1) Depression or depression/ anxiety 2) GP referral/BDI screening 3) Duration 6 mths to 5 yrs 4) >= 14 on BDI	1) Counselling (psychodynamic/ cognitive-behavioural models) 2) No standardisation or manual 3) BACP accredited counsellors 4) 6/12 50-min sessions	Usual GP care (not able to refer to practice counsellor)	1) BDI 2) BSI 3) IIP-32 4) SAS-M 5) Costs	1) Baseline 2) 6 mths 3) 12 mths

Table 4.4 Published Summary Data and Analyses

	Baseline				Short-Term Outcome					
Trial	**Counselling**		**No Counselling**		**Counselling**		**No Counselling**		**Model Summary**	
	N (%)	**Mean (SD)**	**N (%)**	**Mean (SD)**	**N (%)**	**Mean (SD)**	**N (%)**	**Mean (SD)**		
Beck Depression Inventory (BDI)										
Chilvers *et al* 2001	50 (96)	27.1 (8.0)	49 (96)	27.0 (8.0)	39 (75)	15.2 (11.6)	44 (86)	14.8 (10.1)	Unadjusted independent samples t-test on outcome scores	MD=-0.4, 95% CI -4.4 to 5.1, *p*=0.88
Friedli *et al* 1997	70 (100)	19.3 (8.9)	66 (100)	21.8 (9.3)	59 (84)	11.7 (7.7)	51 (77)	15.6 (10.5)	ANCOVA using mean of two follow-up scores adjusting for baseline	*p*=0.10

King *et al* 2000	67 (100)	25.4 (8.6)	67 (100)	26.5 (8.9)	62 (93)	11.5 (7.7)	62 (93)	17.2 (11.9)	General linear model with between-subjects factors for randomised group (3 levels) and site (2 levels) and a within-subjects factor for time (3 levels)	Group: *F*=1.41, *df*=2,191, *p*=0.25;
										Time: *F*=135.90, *df*=2, 190, *p*=0.000;
										Group-by-Time: *F*=3.874, *df*=4, 380, *p*=0.004

Simpson *et al* 2000	92 (100)	21.5 (6.0)	89 (100)	19.9 (5.7)	82 (89)	16.0 (9.3)	79 (89)	16.0 (8.1)	ANCOVA on outcome scores adjusting for baseline	Effect=0.95, 95% CI –3.3 to 1.42, *p*=0.43
Hospital Anxiety and Depression Scale – Depression Subscale (HADS-D)										
Harvey *et al* 1998	99 (89)	-	47 (92)	-	82 (74)	-	39 (76)	-	Unadjusted independent samples t-test on change scores	Effect=-0.7, 95% CI -2.6 to 1.1, *p*=0.43
General Health Questionnaire (GHQ)										
Boot *et al* 1994	124 (100)	15.7 (6.8)	68 (100)	16.6 (6.6)	67 (54)	6.2 (7.0)	41 (60)	10.6 (9.0)	Unadjusted independent samples t-test on outcome scores	*t*=2.82, *df*=106, *p*=0.01
Symptom Index										
Hemmings 1997	136 (94)	1.6 (0.7)	52 (98)	1.5 (0.9)	114 (79)	-	40 (75)	-	MANOVA of the change scores with treatment, practice and gender as factors and initial EPQ score as a covariate	*F*<1

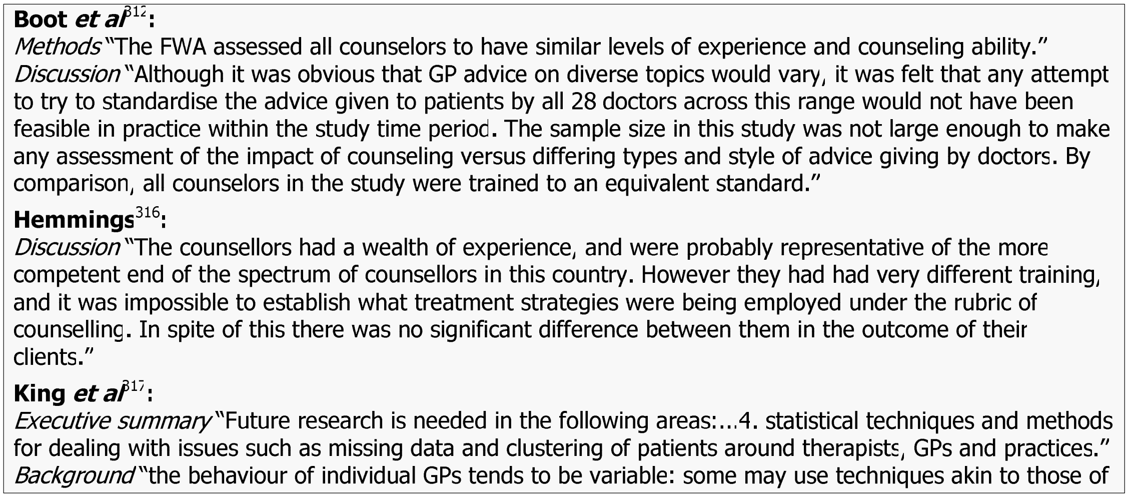
Note: Harvey et al summarised the baseline data as counselling median=10.5, range 1 to 21, no counselling median=12.0, range 2 to 19. The short-term outcomes were summarised as change scores with counselling mean= 2.7 95% CI 1.6 to 3.7, no counselling mean=3.4, 95% CI 1.8 to 4.9; Hemmings summarised the short-term outcomes as adjusted mean change with counselling=0.54 and no counselling=0.52.

one might expect the treatment effect to vary over this timescale, particularly given the treatment duration, this could explain some of the heterogeneity in the treatment effects between studies. The predominant primary outcome was the Beck Depression Inventory[323] (BDI), used within four[315, 316, 319, 320] of the seven trials. Harvey *et al*[317] used the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D), Boot *et al*[314] the General Health Questionnaire (GHQ), and Hemmings[318] the Symptom Index. The GHQ and the Symptom Index encompass more diverse symptom sets.

4.3.2 Published Analyses

The published summary data and analyses that pertain to the Bower and Rowland[37] meta-analysis (see Table 4.2) are given in Table 4.4. No allowance was made in any of the analyses for clustering of outcomes within counsellors. Nor was the extent of the clustering effect reported, or any consideration given to clustering when determining sample size. Despite this, some awareness of the implications of therapist variability was indicated in four of the reports[314, 318-320] (see Box 4.1). Boot *et al*[314] and King *et al*[319] alluded to standardising treatments to remove between-therapist variability. This was achieved by selecting counsellors with similar expertise[314], or by restricting the theoretical model counsellors could apply[319]. Simpson *et al*[320] noted that the treatment delivered varied not only by randomised group but also with the theoretical model, categorising counsellors on this basis, and investigating the differences. Hemmings[318] reported using a preliminary test, but did not give any further details. In each case, there was a sense that more could have been done; Boot *et al*[314] and Simpson *et al*[320] pointing to restrictions of sample size, and King *et al*[319] to the methodology available.

Box 4.1 Reported Handling of Clustering within Studies



Boot *et al*[312]:
Methods "The FWA assessed all counselors to have similar levels of experience and counseling ability."
Discussion "Although it was obvious that GP advice on diverse topics would vary, it was felt that any attempt to try to standardise the advice given to patients by all 28 doctors across this range would not have been feasible in practice within the study time period. The sample size in this study was not large enough to make any assessment of the impact of counseling versus differing types and style of advice giving by doctors. By comparison, all counselors in the study were trained to an equivalent standard."

Hemmings[316]:
Discussion "The counsellors had a wealth of experience, and were probably representative of the more competent end of the spectrum of counsellors in this country. However they had had very different training, and it was impossible to establish what treatment strategies were being employed under the rubric of counselling. In spite of this there was no significant difference between them in the outcome of their clients."

King *et al*[317]:
Executive summary "Future research is needed in the following areas:…4. statistical techniques and methods for dealing with issues such as missing data and clustering of patients around therapists, GPs and practices."
Background "the behaviour of individual GPs tends to be variable: some may use techniques akin to those of

Box 4.1 Continued…

psychological therapists, and the prevalence and quality of antidepressant prescribing may vary widely[322,323] or be influenced by participation in the trial[324]. These variations have important implications for the analysis and interpretation of results" "Psychological therapy practised in primary care is often described as 'eclectic' in nature, using a mixture of theoretical approaches, depending on the clinical context[325,326,327]. However, such therapy is difficult to evaluate because wide variations in the format, process and goals of therapy make it difficult to attribute efficacy to a generic therapy rather than to the particular therapist or therapist–patient relationship. A compromise between internal and external validity necessitates the use of common therapies that are also sufficiently specified to enable them to be reliably distinguished." *Discussion* "Although there were a number of therapists used in the study, the work was not distributed equally, and a small number of therapists were responsible for the management of a significant proportion of the study patients. This factor may make the results dependent on the skill of a small number of therapists. Although the integrity check did indicate that there was sufficient differentiation of therapies and that all the rated CBT sessions were adequate, there was no specific check on the quality of a significant proportion of therapeutic sessions. No formal test was conducted of the possibility of effects associated with the therapists[80]." *Conclusions* "The problems associated with clustering of patients around therapists, GPs and practices have also received significant attention in the methodological literature in recent years[328-330].Future analyses should take account of these clustering effects."

Simpson *et al*[318]:
Executive summary "The results indicated that there were similar improvements for both CBT and psychodynamic counselling, but a larger population may have shown different results." *Background* "There are also a number of other potential problems, including variance in ability between counsellors, the quality of patient–therapist interactions as well as variance among patients. These are documented in more detail elsewhere[331,332]." *Results* "The type of counselling that patients received was not determined by random allocation, but by the model of counselling used by the counsellor attached to a patient's practice…One-way analyses of covariance indicated that there were no significant differences in outcome scores at 6 months between those referred to the psychodynamic counsellors and those referred to the CBT counsellors (*Table 26*). Logistic regression on whether patients were cases or not at 6 months also yielded no significant differences between counselling approaches. However, it must be remembered that the number of patients referred to the CBT counsellors was low, and it would thus be highly unlikely for significant differences to be found."

It is clear from Table 4.4 that a number of the trials were subject to sizeable missing data problems. Last observation carried forward (LOCF) was used in King *et al*[319], with analysis of complete cases used in the other trials. Neither approach is ideal, but for simplicity, and to be consistent with the published meta-analysis, the latter approach was adopted in subsequent chapters. Similarly, five of the trials described departures from the randomised intervention policies[316-320], while Boot *et al*[314] and Chilvers *et al*[315] did not report on this. No further consideration will be given to this within the thesis.

4.4 The Individual-Patient-Data (IPD)

The individual-patient-data relating to the seven trials in the example meta-analysis (Table 4.2) were sought from the trial teams. The Central Office for Research Ethics Committees (COREC) advised that ethical review was not required and that consent from the original researchers would be adequate. The Research and Development (R&D) office covering the South London & Maudsley NHS Trust and the Institute of Psychiatry confirmed that R&D approval would not be required. Consent was given for access to the Boot *et al*[314] data by Pamela Gillies and to the Chilvers *et al*[315] data by

Clair Chilvers. Peter Bower had used the individual-patient-data relating to Friedli *et al*[316], Harvey *et al*[317], King *et al*[319] and Simpson *et al*[320] previously[335] and Chris Roberts had used a subset of the King *et al*[319] data for a presentation[336]. Consent to access these datasets was given by Ian Harvey, Roslyn Corney, Sharon Simpson, Karin Friedli and Michael King. Counsellor identifiers were not available in the Friedli *et al*[316] or Harvey *et al*[317] data held by Peter Bower. They were imputed in the latter, using the GP practice as a proxy, based on a personal communication from Ian Harvey. Karin Friedli suggested that identifiers had been entered, but further communication with Michael King indicated they were no longer available electronically. Approval was given by the Patient Information Advisory Group (PIAG) to access identifiable data to re-enter it in an anonymous form. This was used to create electronic datasets from the paper questionnaires archived by Michael King in relation to Friedli *et al*[316], and by Adrian Hemmings in relation to Hemmings[318].

4.5 Summary of the Illustrative Example

Accordingly, the illustrative example is characterised by the inclusion of randomised trials with partially nested designs. These create an intermediate level in the analysis between patients and studies in the intervention arm only. As none of the original analyses addressed the related clustering or between-arm heteroscedasticity, there is uncertainty over the accuracy of the published meta-analyses. With four of the trials including the Beck Depression Inventory, it was possible to demonstrate methods for combining standardised and absolute mean differences using the same example. The differences in treatment standardisation and the patient populations across studies meant that heterogeneity in the clustering effects was of interest. The main practical issue highlighted was the absence of estimates of the intra-counsellor correlation in all of the original papers, coupled with no details of the preliminary test carried out by Hemmings[318]. The consequences of this are considered in the next chapter.

5 META-ANALYSIS OF INTRACLASS CORRELATION COEFFICIENTS FROM NESTED THERAPIST DESIGNS

5.1 Introduction

Statistical dependence is created amongst the outcomes in randomised trials by cluster sampling of patients or by cluster allocation or delivery of treatments. The implications of clustering associated with therapists for the precision of treatment effect estimates have already been discussed in some detail[286] (see Chapter 2). In brief, the penalties Cornfield[337] described in relation to the cluster randomisation of treatments also apply where cluster allocation of treatments is non-random or multiple randomisations[138] are employed. Additional sampling variation and reduced degrees of freedom are therefore expected. The argument has been made, principally by Crits-Christoph[80, 97], that these penalties can be avoided in the design by "standardizing treatments through the use of treatment manuals and selection, training, certifying, monitoring, and supervising therapists before and during the conduct of an efficacy trial"[97] (p. 520). If the therapist is to be included as a random effect, early advice[79, 80] was to design a trial to include a large number of therapists. More recently, it has been recommended that researchers make appropriate allowance for therapist variation in their sample size calculations[81-84, 96, 97, 286, 336]. If a trial has already been designed and analysed ignoring the potential clustering effects, re-analysis might be carried out with a view to assessing the sensitivity of the conclusions to the presence of clustering. This could be done for a particular trial, or in the context of a systematic review and meta-analysis of treatment effects. The size of the clustering effect may also be of interest in its own right. In each case estimates of the direction and magnitude of the clustering effect are needed which treat therapists as random effects, indexed by intraclass correlation coefficients (ICCs).

Three empirical reviews have been published of the methods used to handle therapist variation in study reports. Martindale[79] included psychotherapy studies published in the *Journal of Abnormal Psychology* during 1973 and 1974, and the *Journal of Consulting and Clinical Psychology* during 1975, but excluded brief and case reports. A decade on, Crits-Christoph and Mintz[80] reviewed comparative studies of psychosocial interventions published between 1980 and February 1990 in the *Journal of Consulting and Clinical Psychology*, excluding 26 (19%) for completely confounding treatment and therapist by involving one therapist per treatment or study. More recently, Lee and Thompson[82] reviewed all individually-randomised trials published in the *British Medical Journal* in 2002. In contrast to previous reviews, they covered a broader range of medical areas and considered all sources of clustering. The principal action noted in all three reviews was for the studies to overlook therapists in analyses. Martindale[79] found that 21 of 33 studies (64%) did so. Crits-Christoph and Mintz[80] reported a similar percentage (68%) for their 114 studies. Of the 17 studies with clustering by care provider located by Lee and Thompson[82], all but one[338] (94%) ignored it in their analysis. Nearly all the other studies carried out preliminary tests[79, 80, 82], consistent with the early advice[79, 80, 97], treating the therapist as a set of fixed effects. It is likely that some preliminary tests also went unreported by study investigators[78]. None of the reviews stated if details of the tests were given, or if the ICC or another proportion-of-variance-explained measure[96, 339-344] was available. Nevertheless, it can be inferred that reporting of ICC estimates in the principal reports of psychotherapy trials is likely to be limited, a recent exception being found in Goodyer *et al*[345]. As a consequence, individual-patient-data (IPD) would ideally be sought to retrospectively estimate the ICC.

A separate literature has developed within the psychotherapy field on therapist effects. This comprises case studies (e.g. Ricks[54] and Strupp[346-349]), observational studies (e.g. Howard *et al*[55], Orlinsky & Howard[56], Brooker & Wiggins[57] and McLellan *et al*[60]), as well as secondary analyses of randomised trials (e.g. Shapiro *et al*[61], Blatt *et al*[350], Project MATCH[351] and Huppert *et al*[352]). The methods of analysis used are diverse, ranging from descriptive to fixed- and random-effects analyses. Measures of therapist effect are similarly diverse, with ICC estimates rarely reported. In 1997, a series of articles were published[50, 353-357] highlighting the therapist as a neglected variable. Lambert and Okiishi[356] suggested that the naturalistic datasets, collected by managed health care companies within the US, provide an opportunity that has been previously unavailable, owing to the much larger numbers of therapists and patients. Since then, Okiishi[62, 358], Wampold and Brown[64], Schoenwald *et al*[359], Baldwin *et al*[67], Lutz *et al*[66], Stiles *et al*[360] and Dinger[68] have analysed large naturalistic samples, contributing to a rising interest in the use of multilevel models in this context. This is echoed in recent re-analyses of the Treatment for Depression Collaborative Research Program trial and the associated commentaries[69-77, 361]. Accordingly, therapist ICC estimates are becoming increasingly available outside principal trial reports, offering a supplementary source when IPD are inaccessible.

A recent review of cluster-randomised trials found that 46% (27 of 59) involved fewer than 10 clusters per arm[362]. The widespread adoption of Crits-Christoph's[80, 97] advice to use standardisation to avoid clustering penalties means that the percentage of nested psychotherapy trials with less than 10 therapists per arm is likely to be higher. This is unfortunate because the precision of ICC estimates is a function of the number of clusters. Non-random allocation of psychotherapies to therapists along with differences in the therapist skills required means that between-treatment heterogeneity is to be anticipated in ICCs. Thus, while the assumption of a common ICC justifies pooling data across arms in cluster-randomised trials, this is more difficult to defend in randomised psychotherapy trials, adding to the imprecision with which ICCs are estimated in this context. One way of alleviating the limitations of single estimates might, therefore, be to pool them across studies.

Blitstein *et al*[38] described a random-effects meta-analytic approach for pooling multiple independent ICC estimates across cluster-randomised studies. This has the advantage of weighting each estimate by its precision incorporating between-study heterogeneity. It fails to consider a number of potential biases in the study estimates however, and to allow for variation in cluster sizes within the studies. It also assumes the within-cluster variance in each study is known. The aim of this chapter was therefore to extend and illustrate methods of obtaining and pooling ICC estimates arising from nested therapist designs. In each case, the proposed methods are contrasted with those used by Crits-Christoph *et al*[363] and Baldwin *et al*[364], and illustrated using the counselling in primary care example introduced in Chapter 4.

5.2 Strategies for Obtaining ICC Estimates

The scarcity of published therapist ICC estimates led Crits-Christoph *et al*[363] and more recently Baldwin *et al*[364] to obtain the IPD relating to a number of psychotherapy trials. Crits-Christoph *et al*[363] wished to summarise and explore predictors of the magnitude of ICC estimates. Baldwin *et al*[364] started a public database, following similar efforts in other settings[271, 365-376]. In both cases the scope was deliberately broad. Crits-Christoph *et al*[363] did not describe the criteria used to select the 15 studies they included; their sample was most probably chosen for reasons of convenience. In contrast, Baldwin *et al*[364] specified their eligibility and search criteria. To be eligible, studies had to involve one or more interventions aimed to reduce an emotional or behavioural problem, each with at least two therapists and two patients per therapist. Searches were of the 2003 and 2004 issues of 8 journals for psychotherapy research. Of 38 studies identified, the authors of 19 (50%) supplied the summary data they requested. One study was then added to give a total of 20 studies. Crits-Christoph *et al*[363] reported summary statistics only, so the best source of published ICC estimates is currently Baldwin *et al*[364]. Their search was limited however, placing restrictions on the completeness of their sample. Researchers wishing to use their database for a specific purpose are, therefore, still faced with issues of missing data.

Heterogeneity is anticipated between ICC estimates for a variety of reasons, statistical and substantive. Numerous potential predictors of ICC estimates have been identified. For cluster-randomised studies, Blitstein *et al*[38] suggested the outcome, its method of measurement, the patient and cluster samples, study design and method of estimation whilst Campbell *et al*[367] considered the study setting, cluster size, type of outcome, its prevalence and method of measurement. In the psychotherapy setting, Crits-Christoph *et al*[363] explored use of a treatment manual, average level of therapist experience, and length and type of treatment. The duration of follow-up may also be a factor. Although many of these predictors are at the study-level, therapist- and patient-level predictors are also possible. Eligibility criteria applied for a specific purpose are therefore likely to be narrower to ensure estimates are matched more closely to the planned or existing study and analysis. If these more restrictive criteria exclude all the published estimates additional assumptions will be required to make use of the database in Baldwin *et al*[364]. Even where some of the estimates do match criteria, researchers are faced with a trade-off between precision and validity.

An alternative approach is to reduce the scope of the eligibility criteria, but extend the search, and obtain IPD in the context of systematic reviews. For counselling in primary care, for example, eligibility criteria for the studies mirror those used in the Cochrane review[37]. As none of the trials included in the review appear in Baldwin *et al*[364], none of their estimates are strictly relevant for this example. As counselling involves the use of eclectic therapeutic approaches for a wide range of social and clinical problems, the study criteria relating to the treated problem and the treatment type might be relaxed. Large naturalistic studies could also be included. Fortunately, the main outcome in the example meta-analysis (Table 4.2) does appear in Baldwin *et al*[364]. As other outcomes do not, outcome criteria could be extended to include all scales measuring depression and general mental health symptoms in the short-term. To justify the assumption that ICC estimates are independent, if multiple outcomes are available for a treatment arm within a study, the closest to the example meta-analysis might be included, excluding outcomes available for multiple informants, e.g. in Couples Therapy. Finally, to ensure the ICC reflects the example meta-analysis, it should have been estimated without any covariate adjustment. Estimates of this kind are "external" to the example and differ in their validity. Conversely, those found "internally", in this case from the IPD, are valid but may be limited in their precision, even when pooled. In the example, two trials[315, 319] used a patient preference design[322], so additional counselling arms were available. Precision could thus be maximised by extending the eligibility in the internal studies or by combining internal and external sources of ICC estimates.

5.3 Methods for Pooling ICC Estimates

5.3.1 Blitstein et al's Random-Effects Meta-Analytic Approach

Blitstein *et al*[38] proposed the following random-effects meta-analysis model for cluster-randomised trials,

$$\hat{\rho}_h = \rho + \varepsilon_h + e_h, \quad h = 1, \ldots, H \qquad (5.1)$$

where $\hat{\rho}_h$ is the ICC estimate observed within study h. Under this model, each study has an associated population ICC ρ_h, which differs from the mean population ICC ρ, by ε_h such that $\varepsilon_h \sim N\left(0, \tau^2_{\{\varepsilon_h\}}\right)$, and from the observed ICC $\hat{\rho}_h$ by e_h such that $e_h \sim N\left(0, \sigma^2_{\{e_h\}}\right)$. Consequently, ε_h denotes random variation at the study-level and e_h the sampling error, with $\{\ \}$ in the subscript referring to the quantity the variance is of. The total variation of $\hat{\rho}_h$ is thus $\mathrm{T}^2_{\{\hat{\rho}_h\}} = \tau^2_{\{\varepsilon_h\}} + \sigma^2_{\{e_h\}}$ where $\tau^2_{\{\varepsilon_h\}}$ and $\sigma^2_{\{e_h\}}$ represent the between- and within-study variances respectively.

In practice, Blitstein *et al*[38] replaced the study estimate $\hat{\rho}_h$ in (5.1) by

$$\hat{\gamma}_h = \ln\left[\frac{1 + (m_h - 1)\hat{\rho}_{A,h}}{1 - \hat{\rho}_{A,h}}\right] = \ln\left[\frac{MSB_h}{MSW_h}\right] = \ln[F_h] \qquad (5.2)$$

where $\hat{\rho}_{A,h}$ is the one-way analysis of variance (ANOVA) estimator given by

$$\hat{\rho}_{A,h} = \frac{MSB_h - MSW_h}{MSB_h + (m_h - 1)MSW_h} \qquad (5.3)$$

MSB_h and MSW_h are the mean squares between and within clusters, m_h is the cluster size, assumed to be equal within studies, and F_h is the *F*-ratio within study h. As Blitstein *et al*[38] pooled transformed study estimates, $\hat{\gamma}_h$, not raw study estimates $\hat{\rho}_h$, their meta-analysis model is more accurately

$$\hat{\gamma}_h = \gamma + \varepsilon_h + e_h, \quad h = 1, \ldots, H \qquad (5.4)$$

where $\varepsilon_h \sim N\left(0, \tau^2_{\{\varepsilon_h\}}\right)$, $e_h \sim N\left(0, \sigma^2_{\{e_h\}}\right)$ and $\mathrm{T}^2_{\{\hat{\gamma}_h\}} = \tau^2_{\{\varepsilon_h\}} + \sigma^2_{\{e_h\}}$.

The rationale Blitstein *et al*[38] gave for using $\hat{\gamma}_h$ in place of $\hat{\rho}_h$ was that, because $\hat{\rho}_h$ is a proportion, its variance is a function of the parameter ρ_h. One consequence of this is that the pooled estimate can be unduly affected by a single study with a small ICC if raw estimates are meta-analysed[371]. Murray *et al*[371] used the sample estimate in place of the population value when estimating the sampling variance. An alternative might have been to use an iterated estimate, replacing the population value by the latest pooled estimate in the second and subsequent iterations. Blitstein *et al*[38] avoided the need for this, using a transformation intended to stabilise the variance across ρ_h. They gave the sampling variance of $\hat{\gamma}_h$, for cluster-randomised trials, as

$$\hat{\sigma}^2_{\{\hat{\gamma}_h\}} = \frac{2}{f_h(k_h - 1)} = \frac{2}{df_{\{SSB\}}} \qquad (5.5)$$

where f_h is the number of trial arms and k_h is the number of clusters per arm, assumed equal across arms within studies, and $df_{\{SSB\}}$ are the degrees of freedom relating to the between-cluster sums of squares.

The transformation in (5.2) closely resembles Fisher's transformation[377], given for one-way ANOVA estimates as

$$\hat{z}_{A,h} = \frac{\hat{\gamma}_h}{2} = \frac{1}{2}\ln\left[\frac{1+(m_h-1)\hat{\rho}_{A,h}}{1-\hat{\rho}_{A,h}}\right] \qquad (5.6)$$

Fisher[377] suggested that the sampling distribution of $\hat{z}_{A,h}$ approaches normality as the number of clusters k_h contributing to the estimate increases, with k_h given irrespective of the number of arms. Fisher[377] gave the approximate sampling variance of $\hat{z}_{A,h}$ as

$$\hat{\sigma}^2_{\{\hat{z}_{A,h}\}} \approx \frac{m_h}{2(m_h-1)(k_h-2)} \qquad (5.7)$$

While alternative approximations have been reported[378, 379], they are also functions of the cluster size, reflecting the imprecision in the within-cluster variance contributing to the ICC. If (k_h-2) in (5.7) is taken to be equal to $df_{\{SSB\}}$ then

$$\hat{\sigma}^2_{\{\hat{\gamma}_h\}} = \hat{\sigma}^2_{\{2\hat{z}_{A,h}\}} = 4\hat{\sigma}^2_{\{\hat{z}_{A,h}\}} \approx \frac{2}{df_{\{SSB\}}}\left(\frac{m_h}{m_h-1}\right) \qquad (5.8)$$

Omitting the factor $(m_h/(m_h-1))$ will have little impact when cluster sizes are extremely large, as is often the case in cluster-randomised studies of communities, for example. However, it notably underestimates the sampling variance (>1%) if cluster sizes are less than 100, and does so by a sizeable amount (>10%) if they are less than 10.

Blitstein *et al*[38] estimated the between-study variance $\tau^2_{\{\varepsilon_h\}}$ using DerSimonian-Laird's[380] (D-L) method of moments estimator

$$\hat{\tau}^2_{\{\varepsilon_h\}} = \max\left\{0, \frac{Q_{\{\hat{\gamma}_h\}}-(H-1)}{\eta_{\{\hat{\gamma}_h\}}}\right\} \qquad (5.9)$$

which assumes the *Q*-statistic has an approximate non-central chi-squared distribution with $H-1$ degrees of freedom and expectation $\eta_{\{\hat{\gamma}_h\}}\tau^2_{\{\varepsilon_h\}}+(H-1)$. In this setting,

$$\hat{Q}_{\{\hat{\gamma}_h\}} = \sum_{h=1}^{H}\hat{\sigma}^{-2}_{\{e_h\}}\left(\hat{\gamma}_h-\bar{\gamma}_h\right)^2 \qquad (5.10)$$

where $\overline{\gamma}_h$ is the arithmetic mean of $\hat{\gamma}_h$ and the variation in the precision of the study estimates between studies is indexed by

$$\hat{\eta}_{\{\hat{\gamma}_h\}} = \sum_{h=1}^{H} \hat{\sigma}^{-2}_{\{e_h\}} - \frac{\sum_{h=1}^{H} \left(\hat{\sigma}^{-2}_{\{e_h\}}\right)^2}{\sum_{h=1}^{H} \hat{\sigma}^{-2}_{\{e_h\}}} \qquad (5.11)$$

It is evident from (5.10) and (5.11) that the D-L between-study variance estimate is a function of the absolute and relative sampling variance estimates $\hat{\sigma}^2_{\{e_h\}}$ and $\eta_{\{\hat{\gamma}_h\}}$, respectively, and of the deviations of the study estimates from their mean, weighted by their estimated precisions, denoted by $Q_{\{\hat{\gamma}_h\}}$. It is thus vulnerable to bias in the estimate of the sampling variance arising from the use of (5.5) when the cluster sizes are small or variable across studies.

Blitstein *et al*[38] gave the pooled estimate of γ as the precision-weighted sum

$$\hat{\gamma} = \sum_{h=1}^{H} \hat{w}_{\{\hat{\gamma}_h\}} \hat{\gamma}_h \qquad (5.12)$$

estimating the study-specific weights and precisions respectively by

$$\hat{w}_{\{\hat{\gamma}_h\}} = \frac{\hat{T}^{-2}_{\{\hat{\gamma}_h\}}}{\sum_{h=1}^{H} \hat{T}^{-2}_{\{\hat{\gamma}_h\}}} \qquad \text{and} \qquad \hat{\mathrm{T}}^{-2}_{\{\hat{\gamma}_h\}} = \left(\hat{\tau}^2_{\{\varepsilon_h\}} + \hat{\sigma}^2_{\{e_h\}}\right)^{-1} \qquad (5.13)$$

They then back-transformed the pooled estimate $\hat{\gamma}$ onto the raw scale using[38]

$$\hat{\rho}_{\{\hat{\gamma}\}} = \frac{\exp\left(\sum_{h=1}^{H} \hat{w}_{\{\hat{\gamma}_h\}} \hat{\gamma}_h\right) - 1}{\exp\left(\sum_{h=1}^{H} \hat{w}_{\{\hat{\gamma}_h\}} \hat{\gamma}_h\right) + \left(\hat{m}_0 - 1\right)} \qquad (5.14)$$

taking variation in the cluster sizes between studies into account using[38]

$$\hat{m}_0 = \overline{m} - \frac{\sum_{h=1}^{H} f_h k_h \left(m_h - \overline{m}\right)^2}{\sum_{h=1}^{H} \left(f_h \left(k_h - 1\right)\right) \sum_{h=1}^{H} \left(\frac{f_h k_h m_h}{f_h k_h}\right)} \quad \text{where} \quad \overline{m} = \frac{\sum_{h=1}^{H} m_h}{H} \qquad (5.15)$$

5.3.2 Biases in Study Estimates

5.3.2.1 Method of Estimation

Although ANOVA estimates of the ICC are frequently used, Fisher[381] originally derived the exact distribution of a pairwise estimator, given as the product-moment correlation averaged over all possible pairs of observations within clusters. This is the maximum-likelihood estimator (MLE) if cluster sizes are equal within studies[382]. Fisher[381] showed that this estimator has a negative bias arising from the "method of calculation" that is independent of the underlying ICC. He[377] gave this for clusters of all sizes as

$$Bias\left(\hat{z}_{ML,h}\right) = -\frac{1}{2}\ln\left(\frac{k_h}{k_h - 1}\right) \quad (5.16)$$

and suggested (pp. 224-5) that it arises because the ratio of the sums of squares, and by extension $\hat{z}_h$, differs by a ratio of $k_h/(k_h - 1)$ from the MLE to the ANOVA estimate. Wang *et al*[383] have elaborated upon this explanation.

Fisher[377] gave the large-sample properties of the ANOVA estimate but did not consider a further bias in this case[379, 383]. One arises because the numerator and denominator of the ICC are not independent so the expectation of the ratio is not equal to the ratio of the expectations[384]. Ginsburg[385] obtained the exact bias, but a simple approximation to this has been derived by Ponzoni and James[384] as

$$Bias\left(\hat{\rho}_{A,h}\right) \approx -\frac{2(1-\rho_h)\left\{\rho_h + \frac{(1-\rho_h)}{m_h}\right\}\left\{\rho_h + \frac{(1-\rho_h)}{k_h m_h}\right\}}{k_h - 1} \quad (5.17)$$

assuming a one-way ANOVA with equal cluster sizes. It is generally negative, tending to zero as the number of clusters increases, but it can be positive if $\rho_h < 0$. In contrast to the previous bias, this is given on the raw scale and depends on the underlying ICC. It is affected by the number of clusters but also by their size. Wang *et al*[383] showed that (5.17) performs well as an approximation with as few as five clusters, except if ρ_h is also between 0.1 and 0.6. They went on to argue that the total bias for the MLE is given on the raw scale as

$$\begin{aligned} Bias(\hat{\rho}_{ML,h}) &= Bias(\hat{\rho}_{A,h}) + (\hat{\rho}_{ML,h} - \hat{\rho}_{A,h}) \\ &= Bias(\hat{\rho}_{A,h}) - \frac{(1-\hat{\rho}_{ML,h})(1+(m_h-1)\hat{\rho}_{ML,h})}{1+m_h(k_h-1)+(m_h-1)\hat{\rho}_{ML,h}} \end{aligned} \qquad (5.18)$$

where $\hat{\rho}_{ML,h} - \hat{\rho}_{A,h}$ on the raw scale precisely equals (5.16) on Fisher's z scale.

It is apparent that both estimators tend to underestimate the population ICC when the number of clusters k_h is small. The extent of this bias varies according to the method of estimation but is consistently larger for the MLE[383, 386]. It is usually less than 10% for ANOVA estimates, even in small samples[386]. Fisher[381] regarded the impact of the bias in (5.16) as relatively unimportant for single estimates, it being of higher order than the standard error, but did see it as a concern when

> "accurate comparisons are made between correlations, and especially averages of correlations, which have perhaps been calculated from samples of different sizes, or by different methods." (Fisher[381] p.235)

Ponzoni and James[384] have expressed a similar view in relation to (5.17). Corrections could be made for these biases by subtracting the relevant bias from the raw estimate. As (5.17) depends on the underlying ICC, the study estimates could be used initially in place of the population parameter, with subsequent iterations substituting the pooled estimate for the population parameter.

5.3.2.2 *Skewed Sampling Distribution*

Fisher's transformation[377, 381] is helpful not only in stabilising the asymptotic variance, but also in normalising the asymptotic sampling distribution, that is otherwise skewed. Konishi[387] showed that Fisher's transformation simultaneously achieves these two aims only when the clusters are of size two. He derived a normalising transformation for the general case, which Konishi and Gupta[388] then used to recommend a modification to Fisher's transformation[377] where the ICC is estimated using maximum-likelihood. They adopted a different but equivalent parameterisation for $\hat{z}_{ML,h}$ in which

$$\hat{z}_{KG,h} = 2\sqrt{\frac{m_h-1}{2m_h}}\hat{z}_{F,h} = \sqrt{\frac{m_h-1}{2m_h}}\ln\left[\frac{1+(m_h-1)\hat{\rho}_{ML,h}}{1-\hat{\rho}_{ML,h}}\right] \qquad (5.19)$$

with variance

$$\hat{\sigma}^2_{\{\hat{z}_{KG,h}\}} = \frac{4(m_h - 1)}{2m_h}\left(\frac{m_h}{2(m_h - 1)(k_h - 2)}\right) = \frac{1}{(k_h - 2)} \qquad (5.20)$$

They[388] gave the bias in the expectation of $\hat{z}_{KG,h}$ as

$$Bias(\hat{z}_{KG,h}) = +\frac{7 - 5m_h}{k_h\sqrt{18m_h(m_h - 1)}} \qquad (5.21)$$

and suggested $\hat{z}_{KG,h} - Bias(\hat{z}_{KG,h})$ as a modified transformation.

It is apparent from (5.21) that the normal approximation of Fisher's transformation[377] becomes less effective, not only as the number of clusters decreases[377], but also as their size increases[387]. As such, the rate at which normality is approached using the classical transformation is less rapid when cluster sizes are greater than two[377]. The effect of the skew is to bias the ICC downward. As the approximate sampling variance Fisher[377] gave depends on the validity of the normality assumption[377], skew will affect its accuracy. In turn, the validity of the chi-squared test based on the Q-statistic, and the estimate of the between-study variance, will also be affected.

If the Konishi-Gupta bias in (5.21) was simply a reflection of the skew in the sampling distribution it would equal zero when $m_h = 2$. It is equal to $-1/2k_h$ however, because Konishi[387, 388] used the MLE in (5.19). The implication of this is that bias arising from the skew is given by

$$Bias(\hat{z}_h) = +\frac{7 - 5m_h}{k_h\sqrt{18m_h(m_h - 1)}} + \frac{1}{2k_h} \qquad (5.22)$$

irrespective of the method of estimation. A correction could be made by subtracting the bias in (5.22) from Fisher's classical transformed study estimate in (5.6). Similarly, the skew bias could be subtracted from the transformed method-corrected raw study estimate, giving a doubly-corrected estimate. The options are summarised in Table 5.1 using Fisher's parameterisation for ANOVA estimates.

Table 5.1 Bias Corrections for ANOVA Study Estimates

Option	Transformed Study Estimate
No Correction	$\frac{1}{2}\ln\left[\frac{1+(m_h-1)\hat{\rho}_{A,h}}{1-\hat{\rho}_{A,h}}\right]$
Method Corrected	$\frac{1}{2}\ln\left[\frac{1+(m_h-1)\hat{\rho}^*_{A,h}}{1-\hat{\rho}^*_{A,h}}\right]$ where $\hat{\rho}^*_{A,h}=\hat{\rho}_{A,h}-(5.17)$
Skew Corrected	$\frac{1}{2}\ln\left[\frac{1+(m_h-1)\hat{\rho}_{A,h}}{1-\hat{\rho}_{A,h}}\right]-\left(\frac{7-5m_h}{k_h\sqrt{18m_h(m_h-1)}}+\frac{1}{2k_h}\right)$
Doubly Corrected	$\frac{1}{2}\ln\left[\frac{1+(m_h-1)\hat{\rho}^*_{A,h}}{1-\hat{\rho}^*_{A,h}}\right]-\left(\frac{7-5m_h}{k_h\sqrt{18m_h(m_h-1)}}+\frac{1}{2k_h}\right)$ where $\hat{\rho}^*_{A,h}=\hat{\rho}_{A,h}-(5.17)$

5.3.2.3 Bounds on Negative ICC Estimates

The range an ICC takes depends on the model in which it is defined[386]. If a variance components model is adopted, the total outcome variance is given as the sum of the between- and within-cluster variances, and the ICC is defined as the proportion of this total that is between clusters,

$$\rho_h=\frac{\sigma^2_{bh}}{\sigma^2_{bh}+\sigma^2_{wh}} \qquad (5.23)$$

As the between-cluster variance cannot be negative, neither can the ICC, so its range lies between zero and one. In the more general "common correlation model"[389, 390], the ICC is specified directly (see Table 5.2). It is the design effect that cannot be negative here, so the lower limit of the ICC falls where the design effect is zero, i.e. where $\rho_h=-1/(m_h-1)$. If clusters are of size two, the range of the ICC is ± 1, but as the cluster size increases the minimum approaches zero. The models are equivalent when $\rho_h\geq 0$, making the choice of model irrelevant[391]. In general, however, the common correlation model is to be preferred because it allows for uncertainty in the direction of the underlying ICC.

Table 5.2 Model Comparison based on a One-Way Analysis of Variance

Source of Variation	Degrees of Freedom	Mean Squares	Expected Mean Squares	
			Under VCM	Under CCM
Between Clusters	k_h-1	MSB_h	$\sigma^2_{wh}+m_h\sigma^2_{bh}$	$\sigma^2_{th}(1+(m_h-1)\rho_h)$
Within Clusters	$k_h(m_h-1)$	MSW_h	σ^2_{wh}	$\sigma^2_{th}(1-\rho_h)$

Note: VCM is the Variance Components Model; CCM is the Common Correlation Model

The range estimates can take depends on the method of estimation[386]. The likelihood of obtaining negative ANOVA estimates due to sampling error is considerable when the population ICC is small and the number of clusters is too[392]. Negative ICC estimates are possible because ANOVA estimation is consistent with common correlation models. Restricting the mean squares to be non-negative is equivalent to restricting the design effect to be so, because the total variance is non-negative by definition. More extreme negative values are to be expected when cluster sizes are small because the minimum ICC estimate varies as a function of the cluster size.

It is common for researchers to omit to report negative estimates or to censor them at zero[393]. One rationale for doing this could be to ensure consistency between the range of the estimates and that of the underlying parameter. Another could be to protect the Type I error rate. Murray *et al*[394] have shown that nominal Type I and II error rates are obtained only when negative ICC estimates are allowed in the analysis. The reason for this can be seen in the simulations reported by Wang *et al*[386]. They compared ANOVA estimates to their exact distributions in small to moderate samples. Omitting negative estimates or censoring them at zero produced upward biases, which were substantial for small population ICCs and small samples. The total bias observed was a trade-off between the upward bias arising from bounding negative estimates and the downward bias arising from the method of estimation. Wang *et al*[386] recommended that negative ICC estimates are reported and incorporated when pooling estimates. As only bounded estimates may be available, a comparison between this recommendation and the inclusion of censored estimates is of interest.

5.3.3 Variability in Cluster Sizes within Studies

The methods described thus far have all assumed the cluster sizes are equal within the studies. This is important because all the transformed study estimates are functions of the ratio MSB_h / MSW_h which is distributed as $F_{k_h-1,\, k_h(m_h-1)}$ if this is the case. Where the clusters vary in size, this distributional property is lost unless $\rho_h = 0$, because MSB_h is no longer distributed as $\chi^2_{k_h-1}$. Extensions are therefore required for $\rho_h \neq 0$.

Wang *et al*[386] suggested replacing m_h by

$$\hat{m}_{0h} = \frac{\sum_{j=1}^{k_h} m_{hj} - \sum_{j=1}^{k_h} m_{hj}^2 \Big/ \sum_{j=1}^{k_h} m_{hj}}{k_h - 1} = \overline{m}_h - \frac{\sum_{j=1}^{k_h} \left(m_{hj} - \overline{m}_h\right)^2}{\left(k_h - 1\right) \sum_{j=1}^{k_h} m_{hj}} \quad \text{where } \overline{m}_h = \frac{\sum_{j=1}^{k_h} m_{hj}}{k_h} \qquad (5.24)$$

in the expectations for MSB_h (Table 5.2) and thus by extension in (5.3), (5.6), (5.7), (5.17) and (5.18). A similar suggestion has been made by Konishi *et al*[395] and Donner and Zou[396] in relation to (5.19) to (5.21). While estimates of the variance components provided using (5.24) are unbiased[397], the ratio of mean squares is only approximately distributed $F_{k_h-1,\, k_h(m_h-1)}$. Thomas and Hultquist[398] proposed an alternative approximation to $F_{k_h-1,\, k_h(m_h-1)}$ in which m_h is replaced by the harmonic mean

$$\hat{m}_{1h} = k_h \Big/ \sum_{j=1}^{k_h} \frac{1}{m_{hj}} \qquad (5.25)$$

and the sample variance of the cluster means replaces MSB_h in the numerator of F. As a variance cannot be negative, this approximation is appropriate only when $\hat{\rho}_h \geq 0$. Both approximations were compared to the exact results[399, 400] by Donner *et al*[401] for unequal family sizes. They found that (5.24) was increasingly unsatisfactory as the ICC increased, the reverse being true for (5.25). Both approximations were less adequate when the degree of imbalance increased.

Baldwin *et al*[364] replaced m_h by the harmonic mean. It is unclear whether MSB_h was also replaced by the sample variance of the cluster means. While exact results would be ideally used[401], $\hat{m}_{0h}$ was substituted here because it involves a simple modification, the ICC is expected to be small and negative ICC estimates are likely. It is of note that the arithmetic mean of the cluster sizes $\overline{m}_h$ will be close to $\hat{m}_{0h}$ unless there is substantial variability in the cluster sizes.

5.4 Application to Counselling in Primary Care

The methods described apply to cluster-randomised trials, where the assumption that the ICC is common to all treatment arms permits pooling across arms within studies. A trivial extension, when between-arm heterogeneity is anticipated, is to restrict the data on an arm-by-arm basis and calculate treatment-specific estimates. As counsellors are involved in the counselling but not the control arms, the methods were applied only to the counselling arms. Consequently, the number of study arms $f_h = 1$ and k_h refers to the number of counsellors per study.

5.4.1 External Study Estimates

The first step was to extract relevant ICC estimates from Baldwin *et al*[364] (see Table 5.3). Of the 343 estimates reported, 322 (94%) were excluded (172 for adjusting for the corresponding baseline, 134 for having dissimilar outcomes, and 16 to avoid there being multiple estimates for arms within studies). The 21 remaining estimates related to 14 psychotherapy studies. The outcome that was most frequently observed was the Beck Depression Inventory (BDI). The Symptom Checklist-90, Brief Symptoms Index, and Life Style Questionnaire are comparable to the Symptom Index. The studies were predominantly small and the number of therapists per arm ranged from 2 to 581. The mean cluster size across arms ranged from 2.5 to 18.4 and the estimates ranged from -0.23 to 0.53. By far the most precise ICC estimate is given by Wampold and Brown[64]. They censored their sample excluding data linked to therapists who saw less than four patients. This apart, it is perhaps the most relevant external estimate for counselling in primary care. This is because the treatments were heterogeneous, as were the treated problems. US outpatient psychotherapy services also broadly resemble the counselling services in primary care in the UK.

5.4.2 Internal Study Estimates

The second step was to obtain ICC estimates relating to the example meta-analysis[37] (see Table 5.4)[*]. To begin with this was accomplished with the sample variances in the published reports and an estimator proposed by Kwong and Higgins[39]. Subsequently, the IPD was used to estimate the ICC directly. Due to the convergence problems that were encountered using ML and restricted ML estimation in this context, ANOVA is the only method reported. As further data was available within the studies, the third step was to obtain ICC estimates for all the available data. These are given in the appendix (see Section 5.6). The ten patients Hemmings[318] had omitted were included, as were

[*] Counsellor IDs were missing for 7 and 11 patients with outcome data for King *et al*[319] and Simpson *et al*[320] respectively. These patients were excluded from all analyses of the IPD.

Table 5.3 Relevant External ICC Estimates taken from Baldwin *et al*[364]

Study	Study Type	N_h	k_{hi}	$\overline{m}_h$	Treated Problem	Manual Used	Outcome	Treatment	ρ	(Residual Error)
Abramowitz *et al*[402]	Efficacy	40	5.0	2.87	Obsessive Compulsive Disorder	Yes	BDI	Intensive Exposure	0.183	(53.16)
							BDI	Twice-Weekly Exposure	0.532	(50.44)
Carlbring *et al*[403]	Efficacy	30	3.0	7.43	Panic Disorder	Yes	BDI	Internet Cognitive Behavior Therapy	-0.137	(75.32)
Ehlers *et al*[404]	Efficacy	28	3.0	8.31	Posttraumatic Stress Disorder	Yes	BDI	Cognitive Therapy	-0.093	(43.39)
Marijuana Treatment Project Research Group[405]	Efficacy	276	12.0	7.04	Cannabis Dependence	Yes	BDI	2-Session Motivational Enhancement Therapy	0.022	(68.81)
							BDI	9-Session Motivational Enhancement Therapy	-0.022	(58.91)
Taylor *et al*[406]	Efficacy	60	2.0	7.75	Posttraumatic Stress Disorder	Yes	BDI	EMDR	0.005	(76.72)
							BDI	Exposure	0.130	(134.11)
							BDI	Relaxation	-0.180	(191.82)
van Minnen *et al*[407]	Efficacy	15	5.0	2.51	Trichotillomania	Yes	BDI	Cognitive Therapy	-0.208	(52.28)
							SCL-90	Behavior Therapy	-0.226	(1259.54)
Watson *et al*[408]	Efficacy	66	7.5	4.25	Depression	Yes	BDI	Cognitive Behavior Therapy	-0.013	(133.30)
							BDI	Process-Experiential Therapy	0.021	(125.74)
Kuyken[409]	Effectiveness	105	20.0	3.32	Depression	Yes	BDI-II	Cognitive Therapy	0.051	(153.78)
Lincoln *et al*[410]	Effectiveness	147	9.5	2.92	Social Phobia	No	BDI	Cognitive Behavior Therapy	-0.133	(81.00)
Merrill *et al*[411]	Effectiveness	186	8.0	17.19	Depression	Yes	BDI	Cognitive Therapy	0.028	(104.55)
Trepka *et al*[412]	Effectiveness	30	6.0	4.18	Depression	Yes	BDI	Cognitive Therapy	0.183	(127.06)
Lange *et al*[413]	Efficacy	69	18.0	2.92	Posttraumatic Stress Disorder	Yes	SCL-90 Depression	Interapy	0.058	(117.86)
Szapocznik *et al*[414]	Efficacy	129	3.0	18.40	HIV-Positive African Americans: Distress, Hassles, Support	Yes	BSI	Structural Ecosystems Therapy	-0.054	(0.61)
							BSI	Person Centered Approach	-0.023	(0.43)
Wampold & Brown[64]	Effectiveness	6146	581.0	9.68	Mixed	No	LSQ	Treatment as Usual	0.078	(288.59)

Note. All studies have nested or partially nested designs. N_h = total sample size; k_{hi} = Mean number of therapists contributing to any given ICC; $\overline{m}_h$ = Mean number of patients per therapist contributing to any given ICC; ρ =ICC post-treatment; BDI = Beck Depression Inventory; SCL-90 = Symptoms Checklist-90; BSI = Brief Symptoms Index; LSQ = Life Style Questionnaire

Table 5.4 Internal ICC Estimates for the Short-Term Mental Health Outcomes reported in Bower & Rowland[37]

Non-Censored Estimates:

	Raw Scale			Transformed Scale				
Study	Kwong & Higgins Estimate	ANOVA Estimate	Method Corrected ANOVA Estimate	Kwong & Higgins Estimate	ANOVA Estimate	Method Corrected ANOVA Estimate	Skew Corrected ANOVA Estimate	Doubly Corrected ANOVA Estimate
Boot 1994	-0.659	-0.029	-0.029	-	-0.217	-0.220	-0.098	-0.101
Chilvers 2001	0.249	0.290	0.308	0.318	0.370	0.394	0.385	0.409
Friedli 1997	-0.842	-0.023	-0.023	-	-0.162	-0.163	-0.015	-0.016
Harvey 1998	-0.220	0.090	0.094	-	0.308	0.320	0.369	0.381
Hemmings 1997	-0.621	-0.022	-0.022	-	-0.774	-0.781	-0.558	-0.564
King 2000	-1.622	-0.140	-0.144	-	-0.296	-0.306	-0.269	-0.279
Simpson 2000	0.275	0.045	0.047	0.733	0.172	0.180	0.241	0.249

Negative Raw Estimates Censored at Zero:

	Raw Scale			Transformed Scale				
Study	Kwong & Higgins Estimate	ANOVA Estimate	Method Corrected ANOVA Estimate	Kwong & Higgins Estimate	ANOVA Estimate	Method Corrected ANOVA Estimate	Skew Corrected ANOVA Estimate	Doubly Corrected ANOVA Estimate
Boot 1994	0.000	0.000	0.001	0.000	0.000	0.004	0.118	0.122
Chilvers 2001	0.249	0.290	0.308	0.318	0.370	0.394	0.385	0.409
Friedli 1997	0.000	0.000	0.001	0.000	0.000	0.007	0.147	0.154
Harvey 1998	0.000	0.090	0.094	0.000	0.308	0.320	0.369	0.381
Hemmings 1997	0.000	0.000	0.000	0.000	0.000	0.005	0.216	0.221
King 2000	0.000	0.000	0.001	0.000	0.000	0.002	0.027	0.029
Simpson 2000	0.275	0.045	0.047	0.733	0.172	0.180	0.241	0.249

Note. The study ICC estimates were used in place of the population parameter in the bias for the method of estimation[384]. The formulae used to implement the bias corrections can be found in Table 5.1. The raw scale refers to the ICC scale; the transformed scale refers to Fisher's classical z scale.

patients in the patient preference arms of Chilvers *et al*[315] and patients that had been randomised between counselling and cognitive behavioural therapy in King *et al*[319]. As data was available for multiple time-points and for other outcomes, this was included as well, although not in the meta-analyses which follow. The counsellors in Chilvers *et al*[315] and King *et al*[319] delivered counselling in all the counselling arms in those studies. The same was true of the psychotherapists and cognitive behavioural therapy for King *et al*[319]. Data in these arms were pooled to avoid creating a spurious rise in the number of clusters and dependencies between the clusters. The effect this had was to increase the cluster sizes in these studies. As this, in turn, affects the range estimates can take, they were brought towards the null. It is apparent from the cluster size averages given in the appendix that variable cluster sizes were an issue for Chilvers *et al*[315] and King *et al*[319].

The estimator suggested by Kwong and Higgins[39] for the clustered arm of a partially nested trial is given by

$$\hat{\rho}_{h1} = \frac{s_{h1}^2 - s_{h0}^2}{s_{h1}^2} \qquad (5.26)$$

where s_{h1}^2 refers to the naïve variance in the clustered arm and s_{h0}^2 to the variance in the non-clustered arm. The conditions under which (5.26) was recommended are very restrictive, however. Kwong and Higgins[39] argued that s_{h0}^2 estimates σ_{wh1}^2, the within-cluster variance in the clustered arm, when a random coefficient model is appropriate (see Model 2.6), and that s_{h1}^2 estimates σ_{th1}^2, the total variance in the clustered arm, when the cluster sizes are approximately equal, the ICC is not large and the number of clusters is not small. The rationale for the latter is that the bias in s_{h1}^2 will be small in these circumstances, as can be seen from its expectation

$$\begin{aligned} E\left[s_{h1}^2\right] &= \frac{E\left[SSB_{h1}\right] + E\left[SSW_{h1}\right]}{n_{h1} - 1} \\ &= \frac{(k_{h1} - 1)(1 + (m_{h1} - 1)\rho_{h1})\sigma_{th1}^2 + k_{h1}(m_{h1} - 1)\sigma_{wh1}^2}{n_{h1} - 1} \qquad (5.27) \\ &= \sigma_{th1}^2\left(1 - \frac{(m_{h1} - 1)\rho_{h1}}{n_{h1} - 1}\right) \end{aligned}$$

with n_{h1} denoting the sample size in the clustered arm. This resembles the expectation of the pooled naïve variance given assuming a random intercept model (Model 2.2) by White and Thomas[42] (p.151) and Hedges[415] (p.156). When the cluster sizes vary, m_{h1} is replaced in (5.27) by

$$\sum_{j=1}^{k_{h1}} m_{h1j}^2 \Big/ n_{h1} \qquad (5.28)$$

where (5.28) is a simple adaptation of Formula (15) in Hedges[416].

On the basis of their simulation study, Roberts and Roberts[84] advised against adopting a random coefficient model for partially nested trials. It is clear from (5.26) that when $s_{wh1}^2 < s_{h0}^2$ the estimate will be biased downwards and when $s_{wh1}^2 > s_{h0}^2$ it will be biased upwards. If s_{wh1}^2 is unknown, the extent of this bias is also. While misspecification of the model clearly affects the magnitude of the estimate, it also affects the limits it can take. These are wider for (5.26) than they are for the underlying ICC, in part because of sampling variation. It is evident from (5.27) that the direction of the bias in $\hat{\sigma}_{th1}^2$ depends on the direction of ρ_{h1}, which is also unknown. Its impact is less conspicuous, having opposite effects in the numerator and denominator of (5.26). While (5.26) has an obvious practical appeal, these biases make it a theoretically unattractive estimator. The other properties of its sampling distribution are also unclear.

The one-way ANOVA estimator in the clustered arm i of study h is given by

$$\hat{\rho}_{A,hi} = \frac{MSB_{hi} - MSW_{hi}}{MSB_{hi} + (\hat{m}_{0hi} - 1)MSW_{hi}} \qquad (5.29)$$

It can be seen in Table 5.4 that the raw estimates were negative in a number of trials, partly accounting for the computational problems experienced. The estimates obtained using Kwong and Higgins'[39] estimator differ considerably from the ANOVA estimates. In five of the seven trials, they were below the theoretical lower limit of $-1/(\hat{m}_{0hi} - 1)$, so that Fisher's transformation cannot be applied. The bias was downward for all but Simpson *et al*[320]. It was least extreme for Chilvers *et al*[315]. These results support the recommendation given by Roberts and Roberts[84] to adopt a two-level heteroscedastic model (Model 2.7) for partially nested trials. They also show that access to the IPD is likely to be necessary to accurately estimate internal ICCs.

The fourth step was to obtain unbiased study estimates to combine on Fisher's z-scale. While this was done for the internal and the external estimates, it is illustrated for the internal estimates relating to the example meta-analysis only (Table 5.4). The negative estimates were censored at zero. Censored and non-censored ANOVA estimates were then corrected for the bias arising from the method of estimation and the skew in the sampling distribution. The former was done on the raw scale. The estimates were then transformed onto Fisher's z scale, and the skew corrected on this scale. In the case of the external estimates, the corrected estimates were also averaged across arms on the z-scale within studies. Only a simple mean was possible, because Baldwin *et al*[364] gave the size and number of clusters averaged across arms.

Bias arising from the method of estimation had little impact on the ANOVA estimates, as was expected. Use of study estimates in place of population values, and of Ponzoni and James'[384] approximation as opposed to the exact bias, leads to some residual bias, but this is not expected to be important[383]. The differences that were observed largely reflect the size and direction of the ANOVA estimates. The impact of this bias in this example would have been comparable for all of the studies if a common ICC had been assumed. Bias arising from the skew had a more discernible effect, particularly where trials had larger cluster sizes. The average cluster size $\hat{m}_{0hi}$ was very small in Chilvers *et al*[315] and King *et al*[319], which is why the bias is less appreciable in these trials. The small number of large clusters in Hemmings[318] is why the bias is most extreme in this trial. Censoring negative estimates at zero decreased the variation in the estimates across studies. The upward bias that resulted partly offset the downward bias from the skew. The doubly-corrected non-censored ANOVA estimates are theoretically the least biased on the *z*-scale. As such, they are the preferred estimates for pooling.

5.4.3 Sampling Variances of the Transformed Study Estimates

The fifth step was to obtain sampling variances for the study estimates on the z-scale. Again, this was done for the internal and the external estimates. It is illustrated for the internal estimates relating to the example meta-analysis in Table 5.5. The transformed scales are not the same in Blitstein *et al*[38] (5.5) and Fisher[377] (5.7). The former is given on Fisher's *z*-scale to allow the comparison to be direct.

Table 5.5 Sampling Variances of the Transformed Study Estimates

	Blitstein *et al*[38]	Fisher[377]	
Study	$\frac{1}{2f_h(k_h-1)}$	$\frac{m_h}{2(m_h-1)(k_h-2)}$	**Ratio**
Boot 1994	0.125	0.181	0.69
Chilvers 2001	0.038	0.066	0.58
Friedli 1997	0.167	0.272	0.61
Harvey 1998	0.063	0.081	0.77
Hemmings 1997	0.250	0.514	0.49
King 2000	0.042	0.063	0.66
Simpson 2000	0.071	0.094	0.76

It is clear from Table 5.5 that Blitstein *et al*s[38] estimator underestimates the sampling variance to differing degrees across the studies, as a function of both the number of clusters and their size. This is important because studies are weighted by the inverse of these estimates. The choice of estimator affects the relative weight of each study, with Blitstein *et al*s[38] estimator giving more weight to Hemmings[318] but less weight to Boot *et al*[314], Harvey *et al*[317] and to Simpson *et al*[320].

5.4.4 Comparison of Pooled Estimates

The final step was to pool the study estimates (Table 5.6). The naïve method used by Crits-Christoph *et al*[363] was to average the ICC estimates within the studies, and then across them. This is compared to the method proposed by Blitstein *et al*[38] and used by Baldwin *et al*[364]. The first modification was to replace the sampling variance suggested by Blitstein *et al*[38] with the approximation suggested by Fisher[377]. Subsequently, study estimates were additionally replaced by their bias-corrected counterparts. Fisher's[377] *z*-scale was used throughout, to again enable the comparisons to be direct. In each case the pooled estimate is given on the transformed scale, together with its standard error and the D-L[380] estimate of between-study heterogeneity. The *Q*-statistic is given with its associated degrees of freedom and *p*-value as well. The *z*-transformation was then inversed, with the respective pooled estimate inserted to give the pooled ICC estimate. This process was repeated for censored and non-censored estimates and for the four sources of ICC estimates.

The pooled ICC estimates vary according to the method used. It can be seen that use

Table 5.6 Pooled ICC Estimates for the Short-Term Mental Health Outcomes reported in Bower & Rowland[37]

	Non-Censored				Censored			
	Transformed Scale			Pooled ICC Estimate	Transformed Scale			Pooled ICC Estimate
Method	**Pooled Estimate**	**Standard Error**	**Between-Study Heterogeneity** $\tau_{\hat{z}}^2$		**Pooled Estimate**	**Standard Error**	**Between-Study Heterogeneity** $\tau_{\hat{z}}^2$	
SOURCE: INTERNAL #1 (MAIN COUNSELLING ARM)								
Naïve	-	-	-	0.030	-	-	-	0.061
Blitstein *et al*	0.000	0.151	0.075 (*Q*(6)=11.93, p=0.06)	0.000	0.173	0.101	0.000 (*Q*(6)=2.81, *p*=0.83)	0.036
Fisher's Transformation	0.038	0.143	0.028 (*Q*(6)=7.49, p=0.28)	0.007	0.172	0.124	0.000 (*Q*(6)=1.79, *p*=0.94)	0.036
Method-Corrected	0.046	0.140	0.023 (*Q*(6)=7.20, p=0.30)	0.009	0.181	0.124	0.000 (*Q*(6)=1.66, *p*=0.95)	0.038
Skew-Corrected	0.106	0.127	0.005 (*Q*(6)=6.24, p=0.40)	0.021	0.230	0.124	0.000 (*Q*(6)=1.37, *p*=0.97)	0.050
Doubly-Corrected	0.112	0.124	0.000 (*Q*(6)=6.02, p=0.42)	0.022	0.239	0.124	0.000 (*Q*(6)=1.33, *p*=0.97)	0.052
SOURCE: INTERNAL #2 (ALL AVAILABLE COUNSELLING ARMS)								
Naïve	-	-	-	0.011	-	-	-	0.029
Blitstein *et al*	0.036	0.107	0.014 (*Q*(6)=7.24, p=0.30)	0.006	0.132	0.093	0.000 (*Q*(6)=1.57, *p*=0.95)	0.023
Fisher's Transformation	0.064	0.107	0.000 (*Q*(6)=5.35, p=0.50)	0.011	0.137	0.107	0.000 (*Q*(6)=1.20, *p*=0.98)	0.024
Method-Corrected	0.068	0.107	0.000 (*Q*(6)=5.05, p=0.54)	0.011	0.142	0.107	0.000 (*Q*(6)=1.11, *p*=0.98)	0.025
Skew-Corrected	0.116	0.107	0.000 (*Q*(6)=3.94, p=0.68)	0.020	0.189	0.107	0.000 (*Q*(6)=0.96, *p*=0.99)	0.035
Doubly-Corrected	0.120	0.107	0.000 (*Q*(6)=3.73, p=0.71)	0.021	0.193	0.107	0.000 (*Q*(6)=0.96, *p*=0.99)	0.036
SOURCE: EXTERNAL (VARIOUS TREATMENTS)								
Naïve	-	-	-	0.009	-	-	-	0.059
Blitstein *et al*	-0.069	0.198	0.466 (*Q*(13)=212.81, p<0.01)	-0.019	0.134	0.095	0.066 (*Q*(13)=41.41, *p*<0.01)	0.042
Fisher's Transformation	-0.049	0.220	0.553 (*Q*(13)=185.45, p<0.01)	-0.014	0.136	0.106	0.074 (*Q*(13)=36.15, *p*<0.01)	0.043
Method-Corrected	-0.033	0.211	0.503 (*Q*(13)=169.69, p<0.01)	-0.009	0.159	0.092	0.043 (*Q*(13)=26.55, *p*=0.01)	0.051
Skew-Corrected	0.034	0.182	0.351 (*Q*(13)=122.49, p<0.01)	0.010	0.279	0.029	0.000 (*Q*(13)=12.90, *p*=0.46)	0.097
Doubly-Corrected	0.050	0.174	0.311 (*Q*(13)=109.92, p<0.01)	0.015	0.282	0.029	0.000 (*Q*(13)=9.03, *p*=0.77)	0.098
SOURCE: COMBINED (INTERNAL #2 PLUS EXTERNAL)								
Naïve	-	-	-	0.010	-	-	-	0.049
Blitstein *et al*	-0.047	0.136	0.304 (*Q*(20)=207.52, p<0.01)	-0.010	0.133	0.070	0.042 (*Q*(20)=46.23, *p*<0.01)	0.032
Fisher's Transformation	-0.026	0.149	0.348 (*Q*(20)=179.63, p<0.01)	-0.006	0.138	0.078	0.044 (*Q*(20)=40.26, *p*<0.01)	0.034
Method-Corrected	-0.012	0.144	0.316 (*Q*(20)=164.96, p<0.01)	-0.003	0.156	0.070	0.027 (*Q*(20)=32.45, *p*=0.04)	0.039
Skew-Corrected	0.059	0.123	0.206 (*Q*(20)=114.26, p<0.01)	0.013	0.273	0.028	0.000 (*Q*(20)=14.79, *p*=0.79)	0.074
Doubly-Corrected	0.072	0.117	0.181 (*Q*(20)=102.85, p<0.01)	0.017	0.276	0.028	0.000 (*Q*(20)=11.57, *p*=0.93)	0.075

Note: The first iteration pooled estimate was used in place of the population parameter in the Ponzoni and James[384] approximation to the bias arising from use of ANOVA as a method of estimation.

of Blitstein *et al*s[38] estimate of the sampling variance of the transformed ICC estimates inflates the *Q*-statistic, and hence the D-L[380] estimate of between-study heterogeneity. To a lesser extent so does bias in the transformed ICC estimates. In this example, the consequence is that the pooled ICC estimate is biased downward. Censoring negative ICC estimates at zero had the reverse effect, upwardly biasing the pooled estimate. This is because of the upward bias in the study estimates and an artificial reduction in between-study heterogeneity. It is of note that there was no heterogeneity between studies, in the internal estimates relating to the example meta-analysis, when the bias was removed. This implies that the range of ICC estimates observed is compatible with sampling variation alone, and that a fixed-effects meta-analysis model may be sufficient.

The non-censored doubly-corrected pooled ICC estimates do not vary much according to the source. When comparing the internal estimates, a trade-off is apparent between the impact of misspecifying the sampling variance and the skew that appears to cancel out in this example. Larger cluster sizes in the available internal data reduce bias from the former but increase it from the latter. Nevertheless, they do increase the precision of the pooled estimates. The naïve estimates can be seen to be both downwardly and upwardly biased, and to vary more according to the source. This is because they take no account of the relative precisions of the estimates or of the shape of their sampling distribution, so are more sensitive to outliers and to the extent of the skew. The clear heterogeneity amongst the external and combined studies is investigated next.

5.4.5 Sources of Between-Study Heterogeneity

The external estimate given by Wampold and Brown[64] could be regarded an outlier, in the sense that treatment standardisation is minimal due to the naturalistic setting, but also because the number of contributing clusters exceeds that for all the other studies combined. At 0.078 it is also almost four times larger than the pooled internal estimate of 0.022. As such, its inclusion is one likely explanation for the heterogeneity apparent, so it was excluded (see Table 5.7). The other obvious reason for heterogeneity is the mixture of outcomes. Since the BDI is the predominant outcome, its ICC is of specific interest. The estimates relating to the BDI are therefore given in Table 5.8. It can be seen in Table 5.7 that excluding Wampold and Brown[64] removed all the heterogeneity beyond that compatible with sampling variation. The pooled external ICC estimate

Table 5.7 Pooled ICC Estimates: Excluding Wampold & Brown[64]

	Non-Censored				Censored			
	Transformed Scale			Pooled ICC Estimate	Transformed Scale			Pooled ICC Estimate
Method	Pooled Estimate	Standard Error	Between-Study Heterogeneity $\tau_{\hat{z}}^2$		Pooled Estimate	Standard Error	Between-Study Heterogeneity $\tau_{\hat{z}}^2$	
SOURCE: EXTERNAL (VARIOUS TREATMENTS)								
Naïve	-	-	-	0.004	-	-	-	0.057
Blitstein *et al*	-0.043	0.100	0.060 (*Q*(12)=24.60, *p*=0.02)	-0.013	0.096	0.065	0.000 (*Q*(12)=3.94, *p*=0.98)	0.031
Fisher's Transformation	-0.015	0.097	0.032 (*Q*(12)=16.70, *p*=0.16)	-0.005	0.094	0.077	0.000 (*Q*(12)=2.61, *p*=1.00)	0.030
Method-Corrected	0.001	0.090	0.019 (*Q*(12)=14.80, *p*=0.25)	0.000	0.109	0.077	0.000 (*Q*(12)=2.86, *p*=1.00)	0.036
Skew-Corrected	0.050	0.084	0.009 (*Q*(12)=13.41, *p*=0.34)	0.016	0.152	0.077	0.000 (*Q*(12)=3.65, *p*=0.99)	0.051
Doubly-Corrected	0.062	0.077	0.000 (*Q*(12)=11.93, *p*=0.45)	0.020	0.167	0.077	0.000 (*Q*(12)=4.39, *p*=0.98)	0.057
SOURCE: COMBINED (INTERNAL #2 PLUS EXTERNAL)								
Naïve	-	-	-	0.006	-	-	-	0.047
Blitstein *et al*	-0.016	0.074	0.040 (*Q*(19)=31.92, *p*=0.03)	-0.003	0.108	0.053	0.000 (*Q*(19)=5.44, *p*=1.00)	0.026
Fisher's Transformation	0.013	0.070	0.014 (*Q*(19)=22.26, *p*=0.27)	0.003	0.109	0.062	0.000 (*Q*(19)=3.75, *p*=1.00)	0.027
Method-Corrected	0.028	0.064	0.003 (*Q*(19)=19.60, *p*=0.42)	0.006	0.119	0.062	0.000 (*Q*(19)=3.83, *p*=1.00)	0.029
Skew-Corrected	0.074	0.062	0.000 (*Q*(19)=17.52, *p*=0.55)	0.018	0.165	0.062	0.000 (*Q*(19)=4.57, *p*=1.00)	0.042
Doubly-Corrected	0.083	0.062	0.000 (*Q*(19)=15.48, *p*=0.69)	0.020	0.175	0.062	0.000 (*Q*(19)=5.20, *p*=1.00)	0.045

Note: The first iteration pooled estimate was used in place of the population parameter in the Ponzoni and James[384] approximation to the bias arising from use of ANOVA as a method of estimation.

Table 5.8 Pooled ICC Estimates: Restricted to the Beck Depression Inventory (BDI)

	Non-Censored				Censored			
	Transformed Scale			Pooled ICC Estimate	Transformed Scale			Pooled ICC Estimate
Method	**Pooled Estimate**	**Standard Error**	**Between-Study Heterogeneity** $\tau^2_{\hat{z}}$		**Pooled Estimate**	**Standard Error**	**Between-Study Heterogeneity** $\tau^2_{\hat{z}}$	
SOURCE: INTERNAL #1 (MAIN COUNSELLING ARM)								
Naïve	-	-	-	0.043	-	-	-	0.084
Blitstein *et al*	0.044	0.179	0.063 (*Q*(3)=6.09, *p*=0.11)	0.015	0.172	0.120	0.000 (*Q*(3)=2.00, *p*=0.57)	0.061
Fisher's Transformation	0.046	0.171	0.025 (*Q*(3)=3.80, *p*=0.28)	0.015	0.163	0.149	0.000 (*Q*(3)=1.20, *p*=0.75)	0.058
Method-Corrected	0.051	0.170	0.024 (*Q*(3)=3.76, *p*=0.29)	0.017	0.174	0.149	0.000 (*Q*(3)=1.15, *p*=0.77)	0.062
Skew-Corrected	0.091	0.167	0.020 (*Q*(3)=3.65, *p*=0.30)	0.031	0.206	0.149	0.000 (*Q*(3)=1.02, *p*=0.80)	0.075
Doubly-Corrected	0.096	0.167	0.020 (*Q*(3)=3.62, *p*=0.31)	0.033	0.217	0.149	0.000 (*Q*(3)=1.01, *p*=0.80)	0.080
SOURCE: INTERNAL #2 (ALL AVAILABLE COUNSELLING ARMS)								
Naïve	-	-	-	0.009	-	-	-	0.028
Blitstein *et al*	0.054	0.107	0.000 (*Q*(3)=2.83, *p*=0.42)	0.014	0.117	0.107	0.000 (*Q*(3)=0.76, *p*=0.86)	0.032
Fisher's Transformation	0.060	0.121	0.000 (*Q*(3)=2.19, *p*=0.53)	0.016	0.121	0.121	0.000 (*Q*(3)=0.59, *p*=0.90)	0.033
Method-Corrected	0.063	0.121	0.000 (*Q*(3)=2.12, *p*=0.55)	0.016	0.125	0.121	0.000 (*Q*(3)=0.55, *p*=0.91)	0.034
Skew-Corrected	0.100	0.121	0.000 (*Q*(3)=1.89, *p*=0.60)	0.027	0.161	0.121	0.000 (*Q*(3)=0.47, *p*=0.92)	0.045
Doubly-Corrected	0.103	0.121	0.000 (*Q*(3)=1.85, *p*=0.60)	0.028	0.165	0.121	0.000 (*Q*(3)=0.47, *p*=0.93)	0.046
SOURCE: EXTERNAL (VARIOUS TREATMENTS)								
Naïve	-	-	-	0.025	-	-	-	0.069
Blitstein *et al*	0.029	0.105	0.040 (*Q*(9)=15.11, *p*=0.09)	0.010	0.112	0.075	0.000 (*Q*(9)=3.70, *p*=0.93)	0.038
Fisher's Transformation	0.055	0.087	0.000 (*Q*(9)=8.89, *p*=0.45)	0.018	0.108	0.087	0.000 (*Q*(9)=2.47, *p*=0.98)	0.037
Method-Corrected	0.069	0.087	0.000 (*Q*(9)=7.19, *p*=0.62)	0.023	0.124	0.087	0.000 (*Q*(9)=2.82, *p*=0.97)	0.043
Skew-Corrected	0.109	0.087	0.000 (*Q*(9)=7.70, *p*=0.56)	0.037	0.167	0.087	0.000 (*Q*(9)=3.48, *p*=0.94)	0.059
Doubly-Corrected	0.122	0.087	0.000 (*Q*(9)=6.48, *p*=0.69)	0.042	0.182	0.087	0.000 (*Q*(9)=4.29, *p*=0.89)	0.066
SOURCE: COMBINED (INTERNAL #2 PLUS EXTERNAL)								
Naïve	-	-	-	0.020	-	-	-	0.057
Blitstein *et al*	0.037	0.076	0.021 (*Q*(13)=17.89, *p*=0.16)	0.011	0.114	0.061	0.000 (*Q*(13)=4.35, *p*=0.99)	0.036
Fisher's Transformation	0.057	0.071	0.000 (*Q*(13)=11.02, *p*=0.61)	0.017	0.113	0.071	0.000 (*Q*(13)=2.95, *p*=1.00)	0.035
Method-Corrected	0.066	0.071	0.000 (*Q*(13)=9.38, *p*=0.74)	0.020	0.124	0.071	0.000 (*Q*(13)=3.21, *p*=1.00)	0.039
Skew-Corrected	0.106	0.071	0.000 (*Q*(13)=9.49, *p*=0.74)	0.033	0.165	0.071	0.000 (*Q*(13)=3.85, *p*=0.99)	0.054
Doubly-Corrected	0.115	0.071	0.000 (*Q*(13)=8.36, *p*=0.82)	0.036	0.176	0.071	0.000 (*Q*(13)=4.61, *p*=0.98)	0.058

Note: The first iteration pooled estimate was used in place of the population parameter in the Ponzoni and James[384] approximation to the bias arising from use of ANOVA as a method of estimation.

0.020 is very close to the pooled internal ICC estimates, and its standard error is smaller due to the larger number of external studies. The pooled combined ICC is identical, and its standard error smaller still. Even so, it is large enough to leave substantial uncertainty regarding the size of the mean population ICC.

It is clear from Table 5.8 that the ICC estimates relating to the BDI were larger than those for other outcomes. As patient responses to broader outcomes can be expected to be more variable than those to more specific ones, the larger ICC seen in relation to the BDI is an intuitive consequence of lower within-cluster variation, but it may simply be a consequence of sampling variation. It is notable that there is some indication of heterogeneity between estimates from the internal studies when they are restricted to the BDI, although this vanishes when other counselling arms available are included. As effort was made to standardise the counselling given in Friedli *et al*[316] and King *et al*[319] but not in Chilvers *et al*[315] or Simpson *et al*[320], the ICC might be expected to be larger for the latter two[315, 320] due to higher between-counsellor variability. The non-censored doubly-corrected pooled ICC estimate was -0.058 for the former two[316, 319] and 0.171 for the latter two[315, 320] using the main counselling arms, -0.027 and 0.071 respectively using available counselling arms, and 0.031 and 0.057 respectively using the combined internal and external BDI estimates.

5.5 Discussion

The current shortage of therapist ICC estimates has the potential to leave researchers guessing their size or using those from other settings[319]. While Baldwin *et al*[364] have created the first public database of therapist ICC estimates, researchers with a specific purpose are still faced with the need to locate missing estimates. It was proposed that psychotherapy researchers continue their effort by obtaining the IPD and pooling ICCs in the context of systematic reviews. This alleviates some of the precision limitations of single estimates while avoiding the validity limitations of external estimates. It reduces the time spent searching for studies and arguably increases the incentive for trialists to make the IPD available. The collaborative group that is formed are then able to design and carry out more definitive trials, while updating the systematic review to assess the sensitivity of its conclusions to the presence of within-study clustering. Within the early phases, the cost implication of involving large numbers of therapists will inevitably lead trialists to search for ways of avoiding the clustering penalties. Standardisation is one option, but it is only partially effective. An alternative would be to explore the use of a Bayesian approach and the more *ad hoc* df* method suggested by Blitstein *et al*[38].

The random-effects meta-analytic approach proposed by Blitstein *et al*[38] was extended to allow for imprecision in the within-cluster variances, unequal cluster sizes within the study arms, and to correct for bias in the study estimates. It was suggested that if ICCs are to be pooled, this is done on Fisher's *z*-scale using non-censored doubly-corrected treatment-specific estimates, with cluster sizes given by $\hat{m}_{0hi}$ and weights given by the inverse of the sum of Fisher's[377] sampling variance and the D-L estimate of between-study heterogeneity. It was suggested that Fisher's z-scale is an appropriate scale for the meta-analysis of ICCs because, unlike the raw ICC scale, it is unaffected by the cluster sizes within a particular study, ensuring the individual estimates are combined on a comparable scale across studies. This requires further support from simulation work. Approximations were used for the bias in the ANOVA estimate and for average cluster sizes, based on the literature[383, 384, 386, 395, 396]. While simulation studies have been reported[383, 401] that suggest these approximations are reasonable in the current context, further work is needed to evaluate the extent of any residual bias in both the study and pooled estimates under conditions that are typical of psychotherapy trials. The adequacy of Fisher's[377] approximate sampling variance and of the D-L estimate also require further evaluation.

Baldwin *et al*[364] commented on the range of ICCs observed in the psychotherapy trials they reported being wider than that in public health or medicine. The counselling in primary care example suggests that this might be simply due to increased sampling variability, rather than to larger population values. The pooled estimate of 0.022 based on seven randomised trials involved 64 counsellors. In contrast, Wampold and Brown's estimate of 0.078, based on a large naturalistic sample, involved 581 therapists. While it may be tempting to interpret this as indicative of a larger therapist *effect* in a clinical setting, the validity of any causal inferences depends on the allocation of therapists to patients being both concealed and random. Causal language is therefore inappropriate unless an appropriate experimental design has been used. Similarly, one might wish to conclude that the therapist ICC is 0.022 in the context of randomised trials. Unless the number of therapists involved in a trial is large, researchers should take account of sampling variation in the ICC when planning future trials or carrying out meta-analyses of existing ones. This could be achieved with a sensitivity analysis, or more formally by adopting a Bayesian approach.

The extension of the CONSORT statement to cluster-randomised trials[417, 418] included a call to routinely report a coefficient of intraclass correlation for each primary outcome, in the form of an ICC or Hayes and Bennett's[419] coefficient of variation. Regrettably, a parallel call was not made in the extension for non-pharmacologic treatment trials[85]. Information was sought on the number of care providers, the cluster size distribution and the case volume for each treatment arm. However, it was not sufficiently explicit that the information supplied in the flow diagram should reflect the analyses reported. Case volume may not reflect trial volume where the care providers are part-time. Both are potentially of interest, but the latter is used to calculate the ICC and design effect. The number of patients allocated to care providers may be different from the number receiving treatment, followed-up, or analysed. The number followed-up and analysed may also change across outcomes and visits. While the summary statistics suggested in the CONSORT extension reflect recognition that the cluster size distribution is likely to be skewed, it would be helpful if the average cluster sizes directly applicable for use by meta-analysts and those performing sample size calculations[115, 420] were reported as well. Consequently, it is recommended that non-censored ANOVA estimates of the ICC are routinely provided in the principal reports of psychotherapy trials, accompanied by the statistical model, number of therapists and average cluster size that relate to these estimates.

5.6 Appendix: Internal Study Estimates using All Available Data

COUNSELLING IN PRIMARY CARE: INDIVIDUAL INTERNAL ESTIMATES POOLING ARMS WITHIN STUDIES

Variable	Visit	Study	Number of Clusters	Cluster Size			ANOVA Estimate of ICC	MSW (Ratio with MSR)
				Arithmetic Mean	$\hat{m}_{0hi}$	Harmonic Mean		
BDI	Baseline	Chilvers 2001	23	7.30	7.09	3.43	-0.034	61.39 (0.97)
		Friedli 1997	4	17.50	15.27	13.28	-0.070	83.23 (0.98)
		King 2000	14	7.79	6.06	2.07	-0.049	84.68 (1.08)
		Simpson 2000	8	8.88	8.77	7.94	0.013	33.61 (1.03)
	Week 8	Chilvers 2001	22	6.50	6.29	3.00	0.068	100.11 (0.99)
	Month 3	Friedli 1997	4	14.75	12.19	10.28	-0.023	60.93 (0.55)
	Month 4	King 2000	14	7.21	5.71	2.07	-0.055	72.89 (0.52)
	Month 6	Simpson 2000	8	8.88	8.77	7.94	0.045	87.74 (1.33)
	Month 9	Friedli 1997	4	15.50	12.74	10.83	-0.025	73.45 (0.64)
	Month 12	Chilvers 2001	22	5.45	5.29	2.65	-0.015	132.91 (0.78)
		King 2000	13	7.00	5.62	2.02	0.015	69.48 (0.97)
		Simpson 2000	8	8.25	8.15	7.43	0.090	92.43 (1.25)
HADS: Depression	Baseline	Harvey 1998	9	11.00	10.39	4.21	0.031	16.87 (0.99)
	Month 4	Harvey 1998	9	9.11	8.64	3.93	0.090	19.28 (0.76)
GHQ	Baseline	Boot 1994	5	21.40	20.11	15.43	0.000	45.58 (1.05)
	Week 6	Boot 1994	5	13.60	12.63	10.05	-0.029	49.56 (0.62)
Symptom Index	Baseline	Hemmings 1997	3	48.33	48.16	47.98	0.017	0.52 (0.70)
	Month 4	Hemmings 1997	3	40.00	39.67	39.39	-0.017	0.41 (0.60)
	Month 8	Hemmings 1997	3	28.33	27.48	26.61	0.016	0.62 (1.00)
BSI: General Severity Index	Baseline	Friedli 1997	4	17.50	15.27	13.28	-0.046	0.5 (1.00)
		King 2000	14	7.64	5.91	2.07	-0.038	0.45 (0.91)
		Simpson 2000	8	8.88	8.77	7.94	0.048	40.97 (0.84)
	Month 3	Friedli 1997	4	14.75	12.19	10.28	-0.047	0.42 (0.58)
	Month 4	King 2000	14	6.93	5.45	2.06	-0.055	0.45 (0.67)
	Month 6	Simpson 2000	8	8.88	8.77	7.94	0.076	91.16 (1.06)
	Month 9	Friedli 1997	4	15.50	12.74	10.83	-0.038	0.42 (0.82)
	Month 12	King 2000	13	6.62	5.33	2.01	-0.126	0.49 (1.04)
		Simpson 2000	8	8.25	8.15	7.43	0.022	126.99 (1.38)

COUNSELLING IN PRIMARY CARE: INDIVIDUAL INTERNAL ESTIMATES POOLING ARMS WITHIN STUDIES – Continued

				Cluster Size				
Variable	**Visit**	**Study**	**Number of Clusters**	**Arithmetic Mean**	$\hat{m}_{0hi}$	**Harmonic Mean**	**ANOVA Estimate of ICC**	**MSW (Ratio with MSR)**
SF-36: General Health	Baseline	Chilvers 2001	23	7.04	6.85	3.41	0.058	437.26 (0.87)
		Harvey 1998	4	10.25	8.80	3.16	0.013	492.22 (1.14)
	Month 4	Harvey 1998	4	8.50	7.24	2.98	0.065	565.94 (0.93)
	Month 12	Chilvers 2001	22	5.32	5.15	2.63	0.195	464.88 (0.61)
SF-36: Physical Functioning	Baseline	Chilvers 2001	23	6.87	6.67	3.38	0.050	479.45 (1.11)
		Harvey 1998	4	10.25	8.80	3.16	-0.068	1082.50 (1.20)
	Month 4	Harvey 1998	4	8.50	7.04	2.88	-0.012	919.46 (1.37)
	Month 12	Chilvers 2001	22	5.23	5.06	2.61	-0.021	646.53 (1.07)
SF-36: Physical Role Limitation	Baseline	Chilvers 2001	23	7.13	6.92	3.41	0.042	1598.10 (1.00)
		Harvey 1998	4	10.00	8.63	3.15	-0.072	1910.99 (1.02)
	Month 4	Harvey 1998	4	8.25	6.65	2.72	-0.057	1909.22 (1.24)
	Month 12	Chilvers 2001	22	5.32	5.16	2.64	0.005	1655.99 (0.91)
SF-36: Bodily Pain	Baseline	Chilvers 2001	23	7.17	6.97	3.45	-0.011	677.14 (0.83)
		Harvey 1998	4	10.00	8.63	3.15	-0.041	877.31 (1.13)
	Month 4	Harvey 1998	4	9.25	7.73	3.01	-0.039	1055.21 (1.51)
	Month 12	Chilvers 2001	22	5.41	5.24	2.65	-0.064	791.48 (0.95)
SF-36: Mental Health	Baseline	Chilvers 2001	23	7.17	6.96	3.45	0.036	219.31 (1.17)
		Harvey 1998	4	10.25	8.80	3.16	0.006	485.20 (1.70)
	Month 4	Harvey 1998	4	9.00	7.57	3.00	0.032	491.95 (1.35)
	Month 12	Chilvers 2001	22	5.36	5.20	2.62	0.064	504.04 (0.70)
SF-36: Vitality	Baseline	Chilvers 2001	23	7.17	6.96	3.45	-0.017	330.20 (1.41)
		Harvey 1998	4	10.25	8.80	3.16	-0.073	301.28 (0.87)
	Month 4	Harvey 1998	4	9.00	7.57	3.00	0.063	416.92 (1.03)
	Month 12	Chilvers 2001	22	5.36	5.19	2.48	0.133	539.26 (0.72)
SF-36: Emotional Role Limitation	Baseline	Chilvers 2001	23	7.09	6.88	3.44	-0.010	899.63 (1.01)
		Harvey 1998	4	10.00	8.63	3.15	0.143	994.47 (1.76)
	Month 4	Harvey 1998	4	8.25	6.87	2.87	-0.061	1726.88 (1.00)
	Month 12	Chilvers 2001	22	5.27	5.11	2.63	0.050	1613.76 (0.82)
SF-36: Social Functioning	Baseline	Chilvers 2001	23	7.22	7.01	3.46	-0.017	566.47 (1.22)
		Harvey 1998	4	10.25	8.80	3.16	-0.045	704.37 (1.24)
	Month 4	Harvey 1998	4	9.25	7.73	3.01	0.074	611.77 (1.47)
	Month 12	Chilvers 2001	22	5.45	5.29	2.65	-0.008	823.38 (0.89)

COUNSELLING IN PRIMARY CARE: INDIVIDUAL INTERNAL ESTIMATES POOLING ARMS WITHIN STUDIES – Continued

Variable	Visit	Study	Number of Clusters	Cluster Size			ANOVA Estimate of ICC	MSW (Ratio with MSR)
				Arithmetic Mean	$\hat{m}_{0hi}$	Harmonic Mean		
SAS-M	Baseline	Friedli 1997	4	17.25	14.95	13.04	-0.024	0.26 (1.17)
		King 2000	14	7.64	5.91	2.07	0.013	0.18 (0.57)
		Simpson 2000	8	8.88	8.77	7.94	0.042	0.17 (0.96)
	Month 3	Friedli 1997	4	14.75	12.19	10.28	-0.029	0.27 (0.90)
	Month 4	King 2000	14	6.79	5.35	2.06	0.007	0.21 (0.48)
	Month 6	Simpson 2000	8	8.88	8.77	7.94	0.097	0.34 (1.41)
	Month 9	Friedli 1997	4	15.50	12.74	10.83	-0.064	0.25 (0.91)
	Month 12	King 2000	13	6.46	5.16	2.01	-0.082	0.26 (0.85)
		Simpson 2000	8	8.25	8.15	7.43	0.080	0.33 (1.02)
IIP-32	Baseline	Hemmings 1997	3	48.33	48.16	47.98	-0.001	0.39 (0.75)
		Simpson 2000	8	8.88	8.77	7.94	0.044	414.43 (1.64)
	Month 4	Hemmings 1997	3	40.00	39.67	39.39	-0.001	0.38 (0.80)
	Month 6	Simpson 2000	8	8.88	8.77	7.94	0.177	360.29 (1.23)
	Month 8	Hemmings 1997	3	28.33	27.48	26.61	0.003	0.38 (1.06)
	Month 12	Simpson 2000	8	8.25	8.15	7.43	0.103	467.62 (1.39)
EPQ: Psychoticism	Baseline	Friedli 1997	4	17.50	15.27	13.28	0.001	12.51 (0.98)
		Hemmings 1997	3	48.33	48.16	47.98	0.038	6.72 (1.04)
EPQ: Extroversion	Baseline	Friedli 1997	4	17.50	15.27	13.28	0.004	29.97 (1.21)
		Hemmings 1997	3	48.33	48.16	47.98	0.021	24.70 (0.77)
EPI: Extroversion	Baseline	Chilvers 2001	23	6.30	6.09	2.85	0.044	22.11 (0.86)
EPQ: Lie	Baseline	Friedli 1997	4	17.50	15.27	13.28	0.034	13.35 (0.89)
		Hemmings 1997	3	48.33	48.16	47.98	0.006	21.12 (1.12)
EPI: Lie	Baseline	Chilvers 2001	23	6.78	6.58	3.33	-0.001	2.55 (0.89)
EPQ: Neuroticism	Baseline	Friedli 1997	4	17.50	15.27	13.28	0.037	25.34 (1.35)
		Hemmings 1997	3	48.33	48.16	47.98	-0.012	15.44 (0.81)
EPI: Neuroticism	Baseline	Chilvers 2001	23	6.35	6.14	2.70	-0.008	27.38 (1.31)
HADS: Anxiety	Baseline	Harvey 1998	9	11.11	10.49	4.22	-0.017	16.58 (1.23)
	Month 4	Harvey 1998	9	9.33	8.85	3.96	-0.020	18.69 (0.92)
Satisfaction	Month 3	Friedli 1997	4	11.50	9.64	7.89	0.010	67.32 (0.54)
	Month 4	King 2000	14	6.86	5.43	2.06	-0.004	0.35 (1.12)
	Month 9	Friedli 1997	4	12.50	10.61	9.22	-0.060	85.20 (0.62)
	Month 12	King 2000	12	6.58	5.32	2.15	0.113	0.43 (0.88)

COGNITIVE BEHAVIOURAL THERAPY IN PRIMARY CARE: INDIVIDUAL INTERNAL ESTIMATES POOLING ARMS WITHIN STUDIES

				Cluster Size				
Variable	Visit	Study	Number of Clusters	Arithmetic Mean	$\hat{m}_{0hi}$	Harmonic Mean	ANOVA Estimate of ICC	MSW (Ratio with MSR)
BDI	Baseline	King 2000	12	9.83	7.37	1.82	0.019	62.56 (0.80)
	Month 4	King 2000	12	8.67	6.49	1.79	-0.060	90.22 (0.64)
	Month 12	King 2000	12	8.00	5.94	1.76	-0.090	116.82 (1.63)
BSI: General Severity Index	Baseline	King 2000	11	10.45	7.89	1.97	0.030	0.43 (0.87)
	Month 4	King 2000	11	8.82	6.67	1.90	-0.010	0.47 (0.69)
	Month 12	King 2000	12	7.42	5.45	1.73	0.012	0.44 (0.93)
SAS-M	Baseline	King 2000	12	9.75	7.33	1.82	-0.010	0.27 (0.85)
	Month 4	King 2000	12	8.08	6.21	1.79	-0.060	0.30 (0.70)
	Month 12	King 2000	11	7.91	5.92	1.88	-0.060	0.27 (0.86)
Satisfaction	Month 4	King 2000	12	8.00	6.10	1.77	-0.070	0.55 (1.75)
	Month 12	King 2000	11	7.18	5.41	1.80	0.000	0.66 (1.35)

6 META-ANALYSIS OF MEAN DIFFERENCES FROM NESTED THERAPIST DESIGNS

6.1 Introduction

At present, there is limited experience of allowing for within-study clustering effects in the meta-analyses of psychotherapy trials (see Chapter 3). However, where allowance has been made for dependence arising from cluster-randomisation, its impact appears to have been minimal[196, 206, 210, 224, 227, 244, 247]. The proportions of meta-analyses and trials that are potentially affected by therapist variation are by comparison higher (see Chapter 3). Studies with therapist trial designs also tend to be smaller, and their ICC estimates less precise, so one might expect the standard error of the pooled treatment effect estimate to be affected to a greater extent by therapist-related clustering effects in the psychotherapy setting. The method by which allowances for clustering are made depends on the original analysis, if published aggregate data are used, but also on the statistical model assumed for studies in the meta-analysis[39, 42, 43, 149-160]. Extensions are needed for fixed- and random-effects meta-analysis models of mean differences that allow for between-arm heteroscedasticity at the therapist and patient levels, consistent with the assumption of a two-level heteroscedastic model[84, 286] for the studies.

Another possible explanation for the limited impact of within-study clustering observed in aggregate meta-analyses is the common practice of ignoring sampling errors arising in the estimation of the study weights. In a frequentist framework, the between-study variance and the sampling variances of study estimates are fixed but unknown. Hardy and Thompson[421] suggested a profile-likelihood approach that allowed for imprecision in the estimation of the between-study variance while still ignoring the sampling errors in the within-study variances, given via the sampling variances of the study estimates. As the between-study variance appears in all study weights in a random-effects meta-analysis, it has been argued that its imprecision is likely to have a larger impact on the standard error of the pooled estimate[40, 421, 422]. Its role in the relative weight each study is given led Whitehead[102] to come to the same conclusion. However, the degrees of freedom available for estimating the variances are also a factor. Viechtbauer[422] showed that it is reasonable to assume the within-study variances are known when the sample sizes are equal across arms, and the average sample size per arm is at least 40. Hardy and Thompson[421] concluded that "Except when all the trials are small, the additional uncertainty would not therefore be expected to have a great impact on the results and so pursuing a full likelihood approach is unnecessarily sophisticated for most practical purposes" (p. 627). On this basis, one would anticipate results from a one-step meta-analysis of the individual-patient-data (IPD) to be almost identical to those using the profile-likelihood approach with the aggregate data.

The presence of within-study clustering and between-arm heteroscedasticity in trials of psychotherapy makes the assumption of known within-study variances less tenable. As more than half the studies in Cochrane reviews involving psychotherapy have sample sizes of less than 50 per arm (see Chapter 3), reductions in their effective sample sizes arising from therapist variation could be important, making the use of a full likelihood approach an attractive option. This is more straightforward to implement if the IPD are available[421]. As this will not always be so, aggregate alternatives are also needed. Sidik and Jonkman[40] have suggested a robust 'sandwich' estimator for the variance of the pooled treatment effect estimate. Although this class of estimator is typically used to handle heteroscedasticity, their purpose was in making allowance for sampling error in the estimated marginal weights of a random-effects meta-analysis. Due to the small number of studies commonly pooled in meta-analyses, they suggested a correction for reducing the associated bias, following Horn *et al*[423] and Royall and Cumberland[424]. It is unclear how effective this would be in the face of within-study clustering, especially when the number of clusters per study is also small.

The presence of between-study heterogeneity in therapist ICC estimates raises further issues. One option is to assume a fully unstructured variance-covariance structure for the random effects, and allow for between-arm heteroscedasticity at the therapist and patient levels on a study-by-study basis. However, the likelihood of negative estimates makes it tricky to fit a one-step meta-analysis of this sort using the IPD. An aggregate approach, in which ICCs are first estimated, analyses of treatment effect ignoring the clustering are then corrected, and these estimates pooled, may therefore be appealing even if the IPD are available. One alternative might be a semi-exchangeable variance-covariance structure where the ICC is assumed to be equal across studies within arms, and treatment-specific ICC estimates are pooled across studies. If the assumption of no between-study heterogeneity in ICCs is acceptable and a one-step meta-analysis of the IPD is used, this represents two meta-analyses for the price of one, reducing the number of steps required. Where it is not, a further option would be to adopt a middle road, and investigate the use of meta-regression models for the random effects.

Focusing initially on treatment effects given in the form of absolute mean differences has several advantages. Firstly, their estimates are unbiased, their sampling variances are independent of the population parameter, and their sampling distribution is exactly normal[422]. This avoids some of the additional complications that are encountered when pooling standardised mean differences or odds ratios and relative risks. It also allows the general implications to be considered before concentrating on those that are more specific. Moreover, the prevalence of continuous outcomes in a psychotherapy setting (see Chapter 3) makes methods for pooling mean differences important in their own right. It also provides a means for analysing multicentre trials of psychotherapy, as the studies in a meta-analysis correspond to the centres in a multicentre trial. The aim of this chapter was therefore to adapt, illustrate and compare methods for pooling mean differences from nested therapist designs. Aggregate methods are described, followed by one-step multilevel models for use with the IPD. In both cases, fixed- and random-effects meta-analysis models will be considered. These are then illustrated using the counselling in primary example introduced in Chapter 4, restricted to the four trials[315, 316, 319, 320] that used the Beck Depression Inventory (BDI) as their primary outcome.

6.2 Aggregate-Data Methods

6.2.1 Standard Fixed- and Random-Effects Meta-Analysis Models

In the simplest meta-analysis model, an underlying treatment effect θ common to all h studies is assumed such that $\theta = \theta_1 = \cdots = \theta_H$. The fixed-effects model implies[102]

$$\hat{\theta}_h = \theta + e_h, \quad h = 1, \ldots, H \qquad (6.1)$$

where $\hat{\theta}_h$ is the treatment effect observed in study h, θ is the population value, and e_h are the sampling errors, with $e_h \sim N\left(0, \sigma^2_{\{e_h\}}\right)$. Heterogeneity in treatment effects observed across studies is therefore ascribed only to sampling error. The more realistic random-effects model permits the population treatment effects to vary across studies, with $\theta_h = \theta + \varepsilon_h$ and $\theta_h \sim N\left(\theta, \tau^2_{\{\varepsilon_h\}}\right)$, where $\tau^2_{\{\varepsilon_h\}}$ is the between-studies variance and θ is now the mean of the population treatment effects. Thus[102]

$$\hat{\theta}_h = \theta + \varepsilon_h + e_h, \quad h = 1, \ldots, H \qquad (6.2)$$

and $\hat{\theta}_h \sim N\left(\theta, \sigma^2_{\{e_h\}} + \tau^2_{\{\varepsilon_h\}}\right)$. The total variance of $\hat{\theta}_h$ is thus $T^2_{\{\hat{\theta}_h\}} = \sigma^2_{\{e_h\}} + \tau^2_{\{\varepsilon_h\}}$, which reduces to a fixed-effects meta-analysis model when $\tau^2_{\{\varepsilon_h\}}$ is zero.

The uniformly minimum-variance unbiased estimate (UMVUE) of the pooled treatment effect θ is given by[146, 147]

$$\hat{\theta}_w = \frac{\sum_{h=1}^{H} w_h \hat{\theta}_h}{\sum_{h=1}^{H} w_h} \qquad (6.3)$$

where $w_h = \dfrac{1}{\sigma^2_{\{e_h\}} + \tau^2_{\{\varepsilon_h\}}}$ is the weight assigned to study h under a random-effects meta-analysis model. Its standard error is given by

$$\sigma_{\{\hat{\theta}_w\}} = \sqrt{\frac{1}{\sum_{h=1}^{H} w_h}} \qquad (6.4)$$

so the two-sided $100(1-\alpha)\%$ confidence interval for $\hat{\theta}_w$ is given by

$$\hat{\theta}_w \pm z_{1-\alpha/2}\sigma_{\{\hat{\theta}_w\}} \qquad (6.5)$$

and the corresponding null hypothesis of no treatment effect is rejected at the$100\alpha\%$ significance level if the test statistic

$$\psi_{\{\hat{\theta}_w\}} = \frac{\left|\hat{\theta}_w\right|}{\sigma_{\{\hat{\theta}_w\}}} > z_{1-\alpha/2} \qquad (6.6)$$

where $z_{1-\alpha/2}$ is the critical value of the standard Normal distribution. It is standard[102] for$\sigma^2_{\{e_h\}}$ and $\tau^2_{\{\varepsilon_h\}}$ to be simply replaced by their respective estimators$\hat{\sigma}^2_{\{e_h\}}$ and $\hat{\tau}^2_{\{\varepsilon_h\}}$.

6.2.2 Sampling Distribution of the Study Mean Differences

Suppose μ_{h1} and μ_{h0} are the true mean outcomes in the intervention and control arm of study h respectively. The population mean difference is then

$$\theta_{MD,h} = \mu_{h1} - \mu_{h0} \qquad (6.7)$$

The outcome of patient l in the ith arm of the hth study is denoted by y_{hil}. If the outcomes can be assumed to be statistically independent both within and across arms, the population variances homogeneous ($\sigma_{h1}^2 = \sigma_{h0}^2 = \sigma_h^2$), and the sample means ($\bar{y}_{h1}$ and $\bar{y}_{h0}$), variances (s_{h1}^2 and s_{h0}^2) and sizes (n_{h1} and n_{h0}) are all available, the study estimate and its sampling distribution are given by[425]

$$\hat{\theta}_{MD,h} = \bar{y}_{h1} - \bar{y}_{h0} \sim N\left(\mu_{h1} - \mu_{h0}, \sigma_h^2\left(\frac{1}{n_{h1}} + \frac{1}{n_{h0}}\right)\right), \quad h = 1, \ldots, H \qquad (6.8)$$

where $\hat{\sigma}^2_{\{\hat{\theta}_{MD,h}\}} = s_h^2\left(\frac{1}{n_{h1}} + \frac{1}{n_{h0}}\right)$ and $s_h^2 = \frac{(n_{h1}-1)s_{h1}^2 + (n_{h0}-1)s_{h0}^2}{n_{h1} + n_{h0} - 2}$

Note that s_h^2 is the analysis of variance estimator of σ_h^2 and that, if the independence and equal variance assumptions hold,

$$\frac{(n_{h1}-1)s_{h1}^2 + (n_{h0}-1)s_{h0}^2}{n_{h1} + n_{h0} - 2} = \frac{SSE_{h1} + SSE_{h0}}{df_{SSE_{h1}} + df_{SSE_{h0}}} = \frac{SSE_h}{df_{SSE_h}} = MSE_h \qquad (6.9)$$

If the outcome variances are heterogeneous across arms (i.e. $\sigma_{h1}^2 \neq \sigma_{h0}^2$) and their ratio is unknown, the study estimate $\hat{\theta}_{MD,h}$ is unaffected but its variance becomes

$$\sigma^2_{\{\hat{\theta}_{MD,h}\}} = \frac{\sigma_{h1}^2}{n_{h1}} + \frac{\sigma_{h0}^2}{n_{h0}} \qquad (6.10)$$

The variances are replaced by s_{h1}^2 and s_{h0}^2 to give the estimator[425] $\hat{\sigma}^2_{\{\hat{\theta}_{MD,h}\}}$. This scenario is referred to as the Behrens-Fisher problem[426]. The analysis of variance estimator s_h^2 is now a linear combination of two independent mean square terms, MSE_{h1} and MSE_{h0}, one for each arm.

Suppose now that the outcome of patient l is nested within the jth cluster of arm i

and is denoted by y_{hijl}. Then assume, for each of h studies, that

i) Two distinct samples of k_{hi} clusters were assigned to intervention and control in a nested therapist trial, implying a two-level heteroscedastic model (2.5);

ii) One sample of $2k_h$ clusters was randomly allocated to intervention or control in a nested therapist trial, implying a random-intercept model (2.2); or

iii) One sample of k_{h1} clusters was assigned to the intervention arm of a partially nested therapist trial, implying a two-level heteroscedastic model (2.7).

Scenarios ii) and iii) can be viewed as special cases of i), where clusters are nested within treatment arms in all three. The study estimate $\hat{\theta}_{MD,h} = \bar{y}_{h1} - \bar{y}_{h0}$ remains a valid estimator of $\theta_{MD,h}$, where the sample means in the clustered arms are given by

$$\bar{y}_{hi} = \frac{\sum_{j=1}^{k_{hi}} \sum_{l=1}^{m_{hi}} y_{hijl}}{\sum_{j=1}^{k_{hi}} m_{hi}} = \frac{\sum_{j=1}^{k_{hi}} \sum_{l=1}^{m_{hi}} y_{hijl}}{n_{hi}} \qquad (6.11)$$

when the cluster sizes are equal within arms.

Analysis of variance estimators of the within- and between-cluster variances, the total and the naïve variance are given in Table 6.1. Simple pooling of the sums of squares across arms is justified by homogeneity of the population within- and between-cluster variances across arms. This assumption is reasonable in the psychotherapy context if a cluster-randomised or multi-tiered design is employed and the source of the clustering is unrelated to treatment delivery. As the psychotherapies are likely to require different skills and performance biases may play a part, between-arm heteroscedasticity may be expected even when psychotherapies are randomly allocated to therapists. However, if homoscedasticity could be assumed, the variance of $\hat{\theta}_{MD,h}$ would be estimated by[107]

$$\hat{\sigma}^2_{\{\hat{\theta}_{MD,h}\}} = s_h^2 \left(\frac{deff_{h1}}{n_{h1}} + \frac{deff_{h0}}{n_{h0}} \right) \qquad (6.12)$$

where $deff_{hi} = \left(1 + (m_{hi} - 1)\rho_h\right)$, $s_h^2 = \dfrac{SSB_h + SSW_h}{df_{\{SSB_h\}} + df_{\{SSW_h\}}}$ and $\rho_h = \dfrac{\sigma_{th}^2 - \sigma_{wh}^2}{\sigma_{th}^2}$

Table 6.1 Analysis of Variance Estimators for Nested Designs

	Between-Arm Heteroscedasticity			Between-Arm Homoscedasticity (Additionally assuming $k_{h1} = k_{h0}$ and $m_{h1} = m_{h0}$)
	Intervention	Control		
		Nested Design	Partially Nested Design	
Within-Cluster Variance	$s_{wh1}^2 = \dfrac{SSW_{h1}}{k_{h1}(m_{h1}-1)}$	$s_{wh0}^2 = \dfrac{SSW_{h0}}{k_{h0}(m_{h0}-1)}$	$s_{h0}^2 = \dfrac{SSE_{h0}}{n_{h0}-1}$	$s_{wh}^2 = \dfrac{\sum_{i=1}^{2} SSW_{hi}}{2k_h(m_h-1)}$
Between-Cluster Variance	$s_{bh1}^2 = \dfrac{SSB_{h1}}{m_{h1}(k_{h1}-1)} - \dfrac{s_{wh1}^2}{m_{h1}}$	$s_{bh0}^2 = \dfrac{SSB_{h0}}{m_{h0}(k_{h0}-1)} - \dfrac{s_{wh0}^2}{m_{h0}}$		$s_{bh}^2 = \dfrac{\sum_{i=1}^{2} SSB_{hi}}{2m_h(k_h-1)} - \dfrac{s_{wh}^2}{m_h}$
Total Variance	$s_{th1}^2 = \dfrac{SSB_{h1}}{m_{h1}(k_{h1}-1)} + \dfrac{SSW_{h1}}{k_{h1}m_{h1}}$	$s_{th0}^2 = \dfrac{SSB_{h0}}{m_{h0}(k_{h0}-1)} + \dfrac{SSW_{h0}}{k_{h0}m_{h0}}$		$s_{th}^2 = \dfrac{\sum_{i=1}^{2} SSB_{hi}}{2m_h(k_h-1)} + \dfrac{\sum_{i=1}^{2} SSW_{hi}}{2k_h m_h}$
Naïve Variance (calculated ignoring clustering)	$s_{h1}^2 = \dfrac{SST_{h1}}{n_{h1}-1} = \dfrac{SSB_{h1} + SSW_{h1}}{k_{h1}m_{h1}-1}$	$s_{h0}^2 = \dfrac{SST_{h0}}{n_{h0}-1} = \dfrac{SSB_{h0} + SSW_{h0}}{k_{h0}m_{h0}-1}$		$s_h^2 = \dfrac{\sum_{i=1}^{2} SST_{hi}}{2(n_h-1)} = \dfrac{(s_{h1}^2 + s_{h0}^2)}{2}$

Note: $SSW_{hi} = \sum_{j=1}^{k_{hi}}\sum_{l=1}^{m_{hi}}(y_{hijl} - \bar{y}_{hij})^2$ $SSB_{hi} = \sum_{j=1}^{k_{hi}} m_{hi}(\bar{y}_{hij} - \bar{y}_{hi})^2$ $SST_{hi} = \sum_{j=1}^{k_{hi}}\sum_{l=1}^{m_{hi}}(y_{hijl} - \bar{y}_{hi})^2$ and $SSE_{hi} = \sum_{l=1}^{n_{hi}}(y_{hil} - \bar{y}_{hi})^2$

so if the cluster sizes are also equal across arms with $m_{h1} = m_{h0} = m_h$, this simplifies to

$$\hat{\sigma}^2_{\{\hat{\theta}_{MD,h}\}} = s_h^2 deff_h \left(\frac{1}{n_{h1}} + \frac{1}{n_{h0}} \right) \qquad (6.13)$$

If a two-level heteroscedastic model[84] applies, pooling of between- or within-cluster sums of squares across arms cannot be justified. In this context, Kwong and Higgins[39] gave the sampling distribution of $\hat{\theta}_{MD,h}$ as

$$\hat{\theta}_{MD,h} = \bar{y}_{h1} - \bar{y}_{h0} \sim N\left(\mu_{h1} - \mu_{h0}, \frac{\sigma^2_{h1} deff_{h1}}{n_{h1}} + \frac{\sigma^2_{h0} deff_{h0}}{n_{h0}} \right), \quad h = 1, \ldots, H \qquad (6.14)$$

where $\hat{\sigma}^2_{\{\hat{\theta}_{MD,h}\}} = \frac{s^2_{h1} deff_{h1}}{n_{h1}} + \frac{s^2_{h0} deff_{h0}}{n_{h0}}$ and $deff_{hi} = \left(1 + \left(m_{hi} - 1\right)\rho_{hi}\right)$

The sampling variance simplifies to[39]

$$\sigma^2_{\{\hat{\theta}_{MD,h}\}} = Var\left[\frac{\sum_{j=1}^{k_{h1}} \sum_{l=1}^{m_{h1}} y_{h1jl}}{n_{h1}} - \frac{\sum_{j=1}^{n_{h0}} y_{h0l}}{n_{h0}} \right] = \frac{\sigma^2_{h1} deff_{h1}}{n_{h1}} + \frac{\sigma^2_{h0}}{n_{h0}} \qquad (6.15)$$

in the case of partial nesting. In a frequentist framework, the population ICC is simply replaced by the relevant internal estimate.

6.2.3 Impact of Sampling Errors in the Estimated Weights

Sidik and Jonkman[40] argue that, if the population weights in (6.3) are replaced by their estimated counterparts, the structure of the variance of the pooled estimate will be of the form

$$\sigma^2_{\{\hat{\theta}_{\hat{w}}\}} = \frac{\sum_{h=1}^{H} \hat{w}_h^2 T^2_{\{\hat{\theta}_h\}}}{\left(\sum_{h=1}^{H} \hat{w}_h \right)^2} \qquad (6.16)$$

When $T^2_{\{\hat{\theta}_h\}} = \hat{w}_h^{-1}$, (6.16) simplifies to the familiar form of this variance, so $\sigma^2_{\{\hat{\theta}_{\hat{w}}\}}$ tends to $\sigma^2_{\{\hat{\theta}_w\}}$ as the sampling errors in the weights decrease. Sidik and Jonkman[40] suggested estimating $T^2_{\hat{\theta}_h}$ with the squared residual $\left(\hat{\theta}_h - \hat{\theta}_{\hat{w}}\right)^2$ to give a 'sandwich' estimator of the variance, adding a correction factor to reduce the bias arising when this is based on a small number of studies. They gave the biased-reduced robust estimator as

$$\hat{\sigma}^2_{\{\hat{\theta}_{\hat{w}}\}} = \frac{\sum_{h=1}^{H} \hat{w}_h^2 \left(1 - \hat{w}_h \Big/ \sum_{h=1}^{H} \hat{w}_h\right)^{-1} \left(\hat{\theta}_h - \hat{\theta}_{\hat{w}}\right)^2}{\left(\sum_{h=1}^{H} \hat{w}_h\right)^2} \qquad (6.17)$$

This is approximately unbiased, providing that $\hat{\sigma}^2_{\{\hat{\theta}_h\}}$ and $\hat{\tau}^2_{\{\theta_h\}}$ are unbiased[40].

The within-study variance estimates given in the previous section are unbiased for the models specified. However, they assume the treatment effect is present in each model so they are biased for fixed-effects meta-analyses[102]. The extent of this bias depends on the size of the study-by-treatment interaction[102]. The most regularly used estimate of $\tau^2_{\theta_h}$ is that of DerSimonian and Laird[380] (D-L),

$$\hat{\tau}^2_{\{\theta_h\}} = \max\left[0, \frac{Q - (H-1)}{\sum_{h=1}^{H} w_h - \left(\sum_{h=1}^{H} w_h^2 / \sum_{h=1}^{H} w_h\right)}\right] \qquad (6.18)$$

where $Q = \sum_{h=1}^{H}\left[\left(\hat{\theta}_h - \hat{\theta}_w\right)^2 / \sigma^2_{\{\hat{\theta}_h\}}\right] = \sum_{h=1}^{H} w_h \left(\hat{\theta}_h - \hat{\theta}_w\right)^2$, and $\hat{\theta}_h$, $\hat{\theta}_w$ and w_h all relate to the corresponding fixed-effects meta-analysis. This is unbiased if the within-study variance is known and homogeneous across studies[422].

Sampling error in the within-study variances means that the test of no between-study heterogeneity is an F test, with $H-1$ and $N-2H$ degrees of freedom, rather than a chi-squared test, with $H-1$ degrees of freedom, under assumptions of independence and common within-study variances[102]. The F statistic is given by $Q/(H-1)$, and N is the total number of patients in a meta-analysis. If there is clustering present within the studies, the denominator degrees of freedom become $K-2H$, where K is the total number of clusters in the meta-analysis. Moreover, if the within-study variances are heterogeneous, the test is no longer an F test, although it can be approximated to one where the denominator degrees of freedom are estimated using a Satterthwaite's procedure[102, 108]. Thus, the D-L estimator is likely to be biased if there is clustering within studies, unless the total number of clusters is large. The additional complication of variable cluster sizes is expected to be of less importance, although it would have a similar effect to heteroscedasticity, in that the test is no longer an F test.

In turn, Sidik and Jonkman's[40] robust `sandwich' estimator is expected to be biased, if it is used in conjunction with the D-L estimator for a random-effects meta-analysis, or study-specific within-study variance estimates for a fixed-effects meta-analysis. One possibility might be use of a robust `sandwich' estimator along with the study-specific within-study variance estimates, as a proxy for a random-effects meta-analysis. This is suggested due to the consistency of the within-study variances with a random-effects meta-analysis and of 'sandwich' estimators in the presence of model misspecification.

6.2.4 Allowance for Finite Samples

In a fixed-effects meta-analysis, the test of no treatment effect is a t test based on $N-H-1$ degrees of freedom, under the assumptions of independence and common within-study variances[102]. In the presence of within-study clustering, the t test is based on $K-H-1$ degrees of freedom, and if the within-study variances differ across arms or studies, it becomes an approximate t test, with the degrees of freedom given by a Satterthwaite procedure[108]. A correction is needed for the t statistic, due to the bias in the within-study variance estimates[102]. In contrast, in a random-effects meta-analysis, the test of no treatment effect is a t test on $H-1$ degrees of freedom, regardless of the random structure at lower levels. As the number of studies contributing to a meta-analysis is finite, Rosner[427] and others[40] have suggested testing hypotheses of the pooled treatment effect using a t as opposed to the usual z statistic. Approximate two-sided $100\alpha\%$ hypothesis tests and $100(1-\alpha)\%$ confidence intervals for $\hat{\theta}_{\hat{w}}$ are thus given in the random-effects meta-analysis by

$$\psi_{\{\hat{\theta}_{\hat{w}}\}} = \frac{\left|\hat{\theta}_{\hat{w}}\right|}{\hat{\sigma}_{\{\hat{\theta}_{\hat{w}}\}}} > t_{H-1,1-\alpha/2} \qquad (6.19)$$

and

$$\hat{\theta}_{\hat{w}} \pm t_{H-1,1-\alpha/2}\hat{\sigma}_{\{\hat{\theta}_{\hat{w}}\}} \qquad (6.20)$$

respectively, where $t_{H-1,1-\alpha/2}$ is the critical value of the t distribution.

6.3 One-Step Multilevel Models of the IPD

6.3.1 Fixed-Effects Meta-Analysis Models

Using notation introduced in Chapter 2 for study-level analyses, where y_l denotes the outcome for the l-th patient, the standard fixed-effects meta-analysis model is[102, 428]

$$y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + e_l \qquad (6.21)$$

where α represents the mean outcome in the control arm of study 1, x_{hl} are indicator variables for the other studies, t_l is an indicator variable for the treatment arm, and β_h and θ are the fixed study and treatment effects respectively. It is commonly assumed that the patient-level residuals e_l are *iid* $N(0, \sigma_e^2)$, though the relaxation of a common patient-level variance across studies has been discussed, so that $e_l \sim N(0, \sigma_{eh}^2)$[102, 428]. It is equally possible to let the patient-level variance vary across arms, in which case the model becomes

$$y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + e_{0l}(1 - t_l) + e_{1l} t_l \qquad (6.22)$$

with the e_{0l} *iid* $N(0, \sigma_{e0}^2)$ and the e_{1l} *iid* $N(0, \sigma_{e1}^2)$. As the assumption of independence is inappropriate due to the presence of therapist variation in this context, model (6.22) can be extended to give the fixed-effects meta-analysis corresponding to model (2.6),

$$y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + u_{0therapist(l)}^{(2)}(1 - t_l) + u_{1therapist(l)}^{(2)} t_l + e_{0l}^{(1)}(1 - t_l) + e_{1l}^{(1)} t_l \quad (6.23)$$

The random effects $u_{0therapist(l)}^{(2)}$ and $u_{1therapist(l)}^{(2)}$ are assumed *iid* $N(0, \sigma_{u0}^2)$ and $N(0, \sigma_{u1}^2)$. In the event that all the studies are partially nested, $u_{0therapist(l)}^{(2)}$ can be constrained to equal zero, and omitted from the model. If only a subset is partially nested, more complex random structures should be considered (see Section 6.3.3). While the study-level random-intercept and random-coefficient models are not recommended for use in the present context, their fixed-effects meta-analysis equivalents can be obtained, for nested designs, by adding fixed study effects to models (2.2) and (2.4) respectively.

6.3.2 Random-Effects Meta-Analysis Models

The standard random-effects meta-analysis is one in which the study effects are fixed but the treatment effect is permitted to vary randomly across studies[102, 428]. If studies have nested therapist designs, the random-effects meta-analysis model becomes

$$y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{0therapist(l)}^{(2)}\left(1 - t_l\right) + u_{1therapist(l)}^{(2)} t_l + e_{0l}^{(1)}\left(1 - t_l\right) + e_{1l}^{(1)} t_l \quad (6.24)$$

where the $\tau_{study(l)}^{(3)}$ are *iid* $N\left(0, \tau^2\right)$ and the random effects are mutually independent. As before, the therapist-level variance in the control arm $u_{0therapist(l)}^{(2)}$ is constrained to equal zero, and the term omitted from the model, if all of the studies are partially nested.

6.3.3 Random-Effects Meta-Regression Models

Meta-regression models have been described that allow the pooled treatment effect to vary subject to one or more study-level characteristics[428-430]. These are used to explore systematic explanations for between-study variation τ^2, and require a large number of studies. Incorporation of a categorical study-level covariate into model (6.24) gives

$$\begin{aligned} y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \sum_{p=2}^{P} W_p s_{pl} t_l + \tau_{study(l)}^{(3)} t_l + u_{0therapist(l)}^{(2)}\left(1 - t_l\right) + u_{1therapist(l)}^{(2)} t_l \\ + e_{0l}^{(1)}\left(1 - t_l\right) + e_{1l}^{(1)} t_l \end{aligned} \quad (6.25)$$

where s_{pl} are indicator variables for the levels of the study characteristic, and W_p are fixed treatment-by-covariate interaction effects. Further categorical or even continuous covariates could be added. When the number of studies is small any variation between studies in the treatment effect is perhaps better left accounted for but unexplained, as in model (6.24). If data are available on one or more therapist-level characteristics, it may be of interest to explore whether the treatment effect varies conditional on these. Here, the covariate varies within studies but is identical for every patient seen by each therapist. Since the number of therapists per study is often small, it may only begin to be feasible to address such questions in a meta-regression. Even so any treatment-by-study-by-covariate interactions may have to be assumed to be zero. As with other IPD meta-regressions, patient-level covariates can also be investigated[428]. In this case, the covariate varies between the patients within the therapists and studies.

Up to this point, the meta-regressions considered are of fixed effects, and in particular of the treatment effect in the form of treatment-by-covariate interactions in a one-step model. Meta-regressions of random effects may also be of interest. The study designs may vary making a more complex variance-covariance structure realistic. The inclusion of fully and partially nested studies is one example. Another is inclusion of studies with and without clustering effects. The standardisation of patient characteristics, or of the therapist's delivery of the treatments, in some studies but not in others is yet another. In these circumstances, there is reason to expect between-study variation in therapist- or patient-level random effects even if there is insufficient statistical power available to detect this. Model (6.24) can be extended for meta-analyses of mixed nested designs as follows,

$$\begin{aligned} y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau^{(3)}_{study(l)} t_l + u^{(2)}_{0therapist(l)}(1-t_l)X_l + u^{(2)}_{1therapist(l)} t_l \\ + e^{(1)}_{0l}(1-t_l)(1-X_l) + e^{(1)}_{0l}(1-t_l)X_l + e^{(1)}_{1l} t_l \end{aligned} \qquad (6.26)$$

where X_l is an indicator variable equal to one when the study has a fully nested design and zero if it is partially nested. Here the residual error in the control arm is allowed to differ across study designs. This ensures that the therapist ICC in the control arm is based on the subset of studies with fully nested designs. It is assumed, as before, that the therapist ICC in the control arm is homogeneous for all fully nested studies. If the assumption of independence is reasonable in some of the studies, model (6.26) can be extended to give

$$\begin{aligned} y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau^{(3)}_{study(l)} t_l + u^{(2)}_{0therapist(l)}(1-t_l)X_l C_l + u^{(2)}_{1therapist(l)} t_l C_l \\ + e^{(1)}_{0l}(1-t_l)(1-C_l) + e^{(1)}_{1l} t_l (1-C_l) + e^{(1)}_{0l}(1-t_l)(1-X_l)C_l + e^{(1)}_{0l}(1-t_l)X_l C_l \\ + e^{(1)}_{1l} t_l C_l \end{aligned} \qquad (6.27)$$

where C_l is an indicator variable equal to one when a study has clustering effects and zero otherwise. Again the residual error is permitted to differ as a function of the study design. For non-clustered studies, it is $e_{0l}^{(1)}(1-t_l)(1-C_l)$ in the control and $e_{1l}^{(1)}t_l(1-C_l)$ in the treatment arm, both terms replaced by $e_l^{(1)}(1-C_l)$ if the patient-level variance is assumed to be homogeneous across arms.

It is reasonable to suppose the patient- and therapist-level variances to be affected by standardising patient or therapist characteristics and behaviour via the use of selection criteria and therapist training, certification, monitoring and supervision. Assuming the study designs are comparable in all other respects, a categorical study-level covariate can be incorporated for the therapist-random effect in model (6.24) as follows,

$$\begin{aligned} y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{0therapist(l)}^{(2)}(1-t_l)T_l + u_{1therapist(l)}^{(2)} t_l T_l \\ + u_{0therapist(l)}^{(2)}(1-t_l)(1-T_l) + u_{1therapist(l)}^{(2)} t_l(1-T_l) + e_{0l}^{(1)}(1-t_l) + e_{1l}^{(1)} t_l \end{aligned} \qquad (6.28)$$

where T_l is an indicator variable equal to one if therapist characteristics or behaviour were standardised and zero otherwise. This might be considered if some of the studies used treatment manuals, while others did not, or if therapists were selected for their expertise, given training, accreditation, monitoring or supervision in some studies but not others. It is assumed in model (6.28) that these design characteristics do not have a simultaneous effect at the patient level. One could instead incorporate a categorical study-level covariate for the patient-level residual error,

$$\begin{aligned} y_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{0therapist(l)}^{(2)}(1-t_l) + u_{1therapist(l)}^{(2)} t_l \\ + e_{0l}^{(1)}(1-t_l)P_l + e_{1l}^{(1)} t_l P_l + e_{0l}^{(1)}(1-t_l)(1-P_l) + e_{1l}^{(1)} t_l(1-P_l) \end{aligned} \qquad (6.29)$$

where P_l is equal to one for studies in which patient characteristics were standardised and zero otherwise. This might be considered if the studies adopt a mix of explanatory and pragmatic approaches to patient eligibility. The potential complexity of the random effects increases with the variability in the study designs. This makes the unstructured alternative appealing but it also reduces the random effects to nuisance parameters. If the number of studies is small, there may be a trade-off between assuming a realistic model for the random effects and computational feasibility too. These models can also be extended to include therapist- and patient-level predictors of the random effects.

6.4 Application to Counselling in Primary Care

Short-term outcomes relating to the BDI were available for 460 patients from four[315, 316, 319, 320] of the counselling in primary care trials. Of these, 224 (49%) were allocated counselling with one of 39 counsellors. Overall, the cluster sizes ranged from 1 to 33, with a median of 3, and an inter-quartile range of 1 to 8. Data were available for 5 or more patients for 18 of the counsellors. Since all four of the trials had a partially nested design, some of the potential complexities were avoided. Given the number of trials, an exchangeable variance-covariance structure was initially assumed for the counselling arm.

6.4.1 Aggregate versus One-Step Meta-Analyses

To reflect common lack of knowledge about the cluster size distribution, equal cluster sizes were assumed for all aggregate analyses. The pooled ICC estimate of 0.033 from Chapter 5, based on the non-censored doubly-corrected internal estimates, was used regardless of the model, and despite the between-study heterogeneity observed. One-step models were implemented in MLwiN using RIGLS, due to its flexibility in modelling the random effects. RIGLS is comparable to REML[95], implemented in xtmixed in Stata. This command has been updated in Version 11 to permit inclusion of one covariate for the patient level error. It can thus be used to implement the simpler models. Details of the programming for both packages are given as an appendix (see Section 6.6).

Table 6.2 summarises the estimates and their standard errors for fixed- and random-effects meta-analyses, progressively relaxing the independence and common variance assumptions within the studies. While the meta-analyses using an aggregate approach all assume an ICC of 0.033, the one-step analyses gave estimates varying between 0.083 and 0.146, depending on the model. The reason for larger one-step estimates is not clear. It may reflect differences in the assumed variance-covariance structure or bias in the RIGLS estimates. Further work is needed to establish the cause. It can be seen that the pooled mean difference and its standard error for a standard aggregate fixed-effects model are -2.429 and 0.886. The associated two-sided 95% CI is -4.17 to -0.69 indicating that counselling reduces short term depression symptoms, measured

Table 6.2 Aggregate versus One-Step Meta-Analyses of the Absolute Mean Difference in BDI between Counselling and Control

AGGREGATE APPROACH

	Ignoring Within-Study Clustering						**Allowing for Within-Study Clustering (Internal #1)**					
Level 1 Variance	**Homogeneous**			**Heterogeneous**			**Homogeneous**			**Heterogeneous**		
	Weights		**Mean Difference**	**Weights**		**Mean Difference**	**Weights**		**Mean Difference**	**Weights**		**Mean Difference**
Study	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**
Chilvers 2001	13.9	19.4	0.36 (2.377)	13.7	19.3	0.36 (2.397)	15.1	19.8	0.36 (2.413)	14.6	19.5	0.36 (2.439)
Friedli 1997	25.7	26.0	-3.86 (1.747)	24.8	25.6	-3.86 (1.785)	23.8	25.0	-3.86 (1.920)	23.9	25.1	-3.86 (1.909)
King 2000	22.9	24.8	-5.75 (1.852)	24.3	25.4	-5.75 (1.803)	24.2	25.2	-5.75 (1.903)	26.0	26.0	-5.75 (1.831)
Simpson 2000	37.6	29.8	-0.45 (1.445)	37.2	29.7	-0.45 (1.458)	37.0	29.9	-0.45 (1.540)	35.5	29.4	-0.45 (1.567)
Fixed: Usual	100	100	-2.429 (0.886)	100	100	-2.475 (0.889)	100	100	-2.424 (0.936)	100	100	-2.526 (0.933)
Fixed: Robust	100	100	-2.429 (1.409)	100	100	-2.475 (1.439)	100	100	-2.424 (1.435)	100	100	-2.526 (1.459)
Random: Usual	100	100	-2.497 (1.402)	100	100	-2.517 (1.415)	100	100	-2.484 (1.417)	100	100	-2.528 (1.427)
Random: Robust	100	100	-2.497 (1.395)	100	100	-2.517 (1.408)	100	100	-2.484 (1.411)	100	100	-2.528 (1.422)
D-L $\hat{\tau}^2_{\theta_h}$	4.500			4.626			4.335			4.480		

ONE-STEP APPROACH

	Ignoring Within-Study Clustering				**Allowing for Within-Study Clustering (Internal #1)**			
Level 1 Variance	**Homogeneous**		**Heterogeneous**		**Homogeneous**		**Heterogeneous**	
Meta-Analysis	**Fixed**	**Random**	**Fixed**	**Random**	**Fixed**	**Random**	**Fixed**	**Random**
Intercept	16.15 (1.139)	15.46 (1.311)	16.31 (1.154)	15.53 (1.365)	15.70 (1.190)	15.36 (1.293)	15.75 (1.245)	15.41 (1.361)
Friedli 1997	-1.29 (1.402)	-0.16 (1.738)	-1.53 (1.391)	-0.26 (1.802)	-0.49 (1.581)	0.00 (1.730)	-0.54 (1.659)	-0.06 (1.820)
King 2000	-0.61 (1.382)	1.01 (1.681)	-0.96 (1.376)	0.87 (1.744)	0.25 (1.490)	1.11 (1.658)	0.09 (1.544)	0.97 (1.738)
Simpson 2000	0.77 (1.318)	0.83 (1.613)	0.72 (1.312)	0.79 (1.675)	0.91 (1.424)	0.93 (1.590)	0.93 (1.484)	0.94 (1.669)
Counselling	-2.469 (0.900)	-2.469 (1.423)	-2.465 (0.897)	-2.467 (1.423)	-2.429 (1.082)	-2.485 (1.446)	-2.458 (1.116)	-2.507 (1.452)
$\hat{\tau}^2$ (Level 3: Int.)		4.802 (4.581)		4.830 (4.462)		3.845 (4.888)		3.560 (4.758)
$\hat{\sigma}^2_v$ (Level 2: Int.)					9.659 (6.044)	8.062 (6.094)	12.531 (6.411)	11.203 (6.540)
$\hat{\sigma}^2_e$ (Level 1: Cont.)	92.773 (6.117)	91.877 (6.085)	102.913 (9.472)	101.878 (9.379)	89.100 (6.069)	88.979 (6.061)	102.198 (9.408)	101.865 (9.377)
$\hat{\sigma}^2_\xi$ (Level 1: Int.)			82.111 (7.759)	81.329 (7.754)			73.198 (7.451)	73.241 (7.457)
Counsellor ICC	-	-	-	-	0.098	0.083	0.146	0.133
-2 Log Likelihood	3384.30	3385.26	3381.43	3382.71	3382.24	3382.86	3376.94	3377.68

by the BDI, by an average of approximately 2.4 points, and that this reduction is statistically significant at the 5% level. The equivalent one-step estimate and its standard error are -2.469 and 0.900 with the two-sided 95% CI based on the t value, -4.24 to -0.70. The similarity of these results implies that bias and sampling error in the aggregate within-study variance estimates is not important under this model in this example. The pooled mean difference and its standard error in the analogous aggregate random-effects model are -2.497 and 1.402. The increase in the standard error arises from between-study heterogeneity in the mean differences across studies. This widens the two-sided 95% CI to -5.24 to 0.25, so the reduction in BDI is no longer statistically significant at the 5% level. If a one-step model had been used, the estimate and its standard error would be -2.469 and 1.423, and the 95% CI using the t value -7.00 to 2.06. The disparity in the standard errors is partly explained by that of the between-study variance estimates (4.500 vs. 4.802), which in turn is due either to bias arising from sampling error or heterogeneity in the within-study variance estimates. Its impact is less pronounced than that of allowing for the finite sampling of studies in the CI in this example. Either way, the evidence in favour of counselling in primary care is less clear if between-study heterogeneity is taken into consideration.

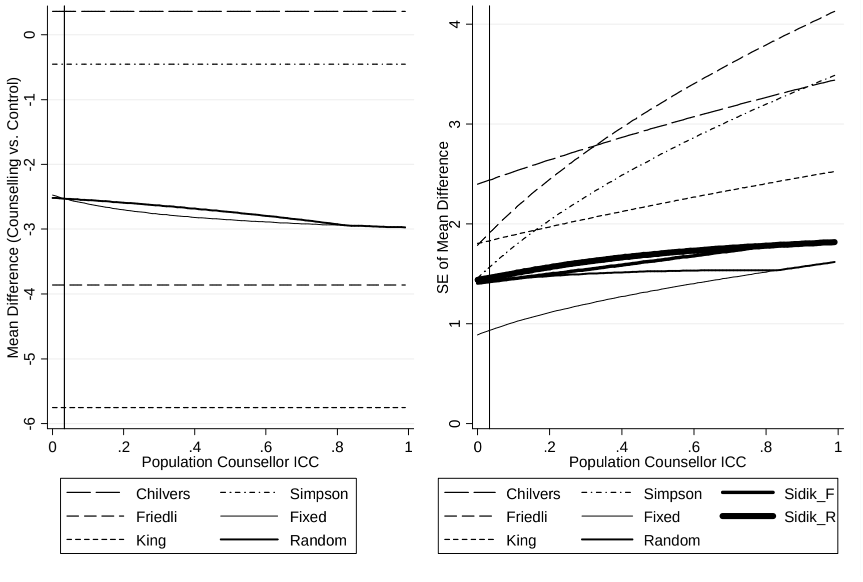
The impact of allowing for between-arm heteroscedasticity and within-study clustering appears to be minimal if the pooled estimates and their standard errors are compared across the usual aggregate analyses. Within-study clustering has a greater impact than between-arm heteroscedasticity in this case, although the effect of the latter is more perceptible in random-effects analyses. A similar pattern can be seen for the one-step analyses. The impact is slightly more pronounced, increasing the disparity between the aggregate and one-step results as the model becomes more realistic. It is of note that the D-L and one-step estimates of between-study heterogeneity differ with the pattern reversing when clustering is taken into account. One-step estimates of the counsellor ICC are also larger than the pooled aggregate estimate (see Chapter 5), and vary from model to model. These differences arise, in part, because the variances are estimated simultaneously in a one-step model, and make appropriate allowance for all the other effects in the model. The exchangeability assumption was not made when pooling ICC estimates in Chapter 5, as this would have required a fixed-effects model for the ICCs. As such, one might expect the disparity between aggregate and one-step results to be greater the larger the between-study heterogeneity in the ICC estimates. In this case, the results continue to be dominated by between-study heterogeneity in the treatment effects.

The most realistic one-step pooled mean difference and standard error are -2.507 and 1.452, with the two-sided 95% CI given by -7.13 to 2.11. The aggregate proxy to this is the corresponding robust fixed effects analysis, where the estimate and its standard error are -2.526 and 1.459, and the two-sided 95% CI is -7.17 to 2.12. Both are very similar, and in each case the confidence interval is marginally wider than the standard random effects one with the conclusion remaining unchanged. What is more striking is a comparison of the robust aggregate results to the usual ones. In the random-effects case, the robust standard errors are smaller than the usual ones. At the same time the robust standard errors are larger for the fixed- than the random-effects analyses. This suggests that bias in the robust estimator is an important consideration, and that care should be taken when using Sidik and Jonkman's[40] estimator in this setting. However, it does show potential for providing an adequate proxy for the one-step results if used in conjunction with a fixed effects analysis.

6.4.2 Sensitivity to the Population ICC

The sensitivity of the mean difference and its standard error to the population ICC are plotted in Figure 6.1. The dashed lines represent the study-level results, while the solid lines correspond to the pooled results for aggregate models allowing for clustering and between-arm heteroscedasticity within the studies.

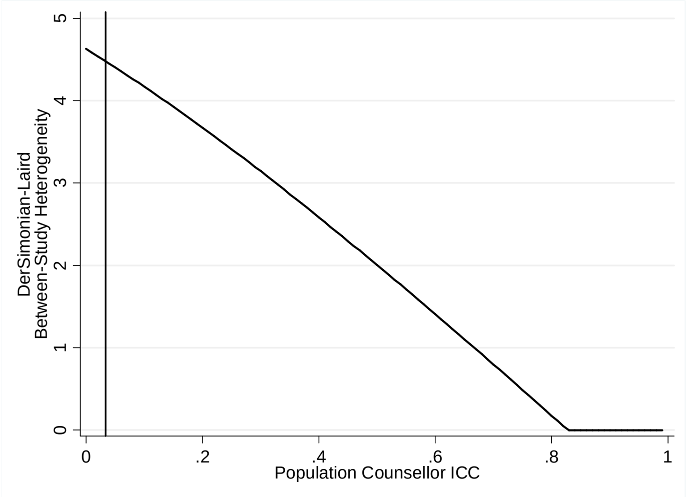
Figure 6.1 Sensitivity of the Mean Difference and its SE to the Population ICC



The study estimates are unaffected as the ICC increases, but the pooled estimates are more extreme. This is because King *et al*[319] has more weight as the ICC increases. This is slightly more pronounced for the fixed-effects estimate. There is some evidence of between-study heterogeneity in the slope of the standard error over the range of the ICC. This questions the validity of the exchangeability assumption in this example. The slope of the pooled standard error is not especially steep, indicating that the results are not sensitive to the ICC in the anticipated range, if between-study heterogeneity in the treatment effects is taken into account.

The relationship between the population ICC and between-study heterogeneity in the treatment effects is depicted below in Figure 6.2. It can be seen that the D-L estimate decreases as the ICC increases, and is censored at zero when the ICC is over 0.8. This implies that the heterogeneity in the mean differences across studies is contributed to, but is not simply explained by, heterogeneity between the counsellors.

Figure 6.2 Sensitivity of Between-Study Heterogeneity to the Population ICC



6.4.3 Meta-Regression Analyses

While the studies all had partially nested designs, Friedli *et al*[316] and King *et al*[319] used a treatment manual and additional training or monitoring to standardise the delivery of counselling. Chilvers *et al*[315] and Simpson *et al*[320], instead, took a pragmatic approach. Patient eligibility was restricted to depression, or comorbid depression and anxiety, in Chilvers *et al*[315], King *et al*[319] and Simpson *et al*[320]. Friedli *et al*[316], in contrast, accepted a broad set of referrals. None of the four studies adopted a doubly pragmatic design.

Table 6.3 summarises the results of two meta-regression models, one for the therapist random effect (model 6.28) and the other for the patient residual (model 6.29). Both explore sources of heterogeneity in the counsellor ICC between studies. Computational problems were avoided by illustrating the extension of the most realistic fixed-effects meta-analysis. If further studies had been available, a random-effects meta-regression would have been preferable.

Table 6.3 Meta-Regression Analyses of the Mean Difference in BDI

	Source of Heterogeneity in ICCs	
Model	**Patient Eligibility**	**Treatment Standardisation**
Intercept	15.73 (1.164)	15.84 (1.268)
Friedli 1997	-0.56 (1.540)	-0.58 (1.375)
King 2000	-0.50 (1.513)	-0.15 (1.452)
Simpson 2000	0.90 (1.387)	0.86 (1.591)
Counselling	-2.373 (1.050)	-3.576 (0.905)
$\hat{\sigma}^2_v$ (Level 2: Int.)	8.624 (5.217)	
$\hat{\sigma}^2_{\xi 0}$ (Level 1: Int./Mix.)	80.712 (9.307)	
$\hat{\sigma}^2_{\xi 1}$ (Level 1: Int./Dep.)	53.318 (10.924)	
$\hat{\sigma}^2_{e0}$ (Level 1: Cont./Mix.)	86.914 (9.318)	
$\hat{\sigma}^2_{e1}$ (Level 1: Cont./Dep.)	142.062 (25.518)	
$\hat{\sigma}^2_{v0}$ (Level 2: Int./Not Stand.)		28.188 (14.074)
$\hat{\sigma}^2_{v01}$ (Level 2 Covariance)		-15.138 (7.026)
$\hat{\sigma}^2_{\xi}$ (Level 1: Int.)		71.708 (7.120)
$\hat{\sigma}^2_e$ (Level 1: Cont.)		102.133 (9.403)
Counsellor ICC (Mixed)	0.097	
Counsellor ICC (Depression)	0.139	
Counsellor ICC (Not Standardised)		0.282
Counsellor ICC (Standardised)		-0.030
-2 Log Likelihood	3368.22	3364.33

A reduction of 8.72 was seen in the log likelihood by including separate residual terms for studies with broad and narrow referrals. The pooled treatment effect reduced very slightly, as did its standard error. As one would expect, the counsellor ICC was higher when the patients were more homogeneous. When distinct therapist-level terms were included for studies standardising counselling and those that did not, the log likelihood reduced by 12.61. Here, the pooled treatment effect also increased appreciably, which reflects the association between the study estimate and counsellor ICC. The standard error also reduced perceptibly, being similar to the standard fixed effects equivalent. As the pooled counsellor ICC is negative for the studies that standardised counselling, a different parameterisation of the model was used which included a covariance term rather than an explicitly negative estimate. Again, as one would expect, the counsellor

ICC was higher when counselling was not standardised. The standard errors for these variance estimates are very large due to the number of studies and counsellors. It was also not possible computationally to simultaneously allow for heterogeneity from both these sources, and to fit the model of choice. The potential to do so when the number of studies available is larger is clear, however. The facility to disentangle the predictors of the components of an ICC is also extremely attractive, as it allows for the possibility that the predictors differ between the components.

6.5 Discussion

Extensions have been described that allow for between-arm heteroscedasticity at the therapist- and patient-levels in meta-analyses of studies with nested therapist designs. Aggregate and one-step models were contrasted, and the potential for exploring meta-regression models for fixed and random effects outlined. The example of counselling in primary care was used to illustrate a selection of the issues that arise in this context. It was shown that the robust `sandwich' estimator of the variance of a pooled treatment effect put forward by Sidik and Jonkman[40] is a promising proxy for a one-step random-effects meta-analysis using the full-likelihood, when used in an aggregate fixed-effects meta-analysis with study-specific within-study variance estimates. A simulation study is needed to investigate this further, and to evaluate the scenarios under which sampling errors in the estimated weights can be safely ignored, or a profile-likelihood approach adopted. A Bayesian approach would be of interest in this respect, to fully account for sampling error in the ICC, and to formally incorporate external estimates of the ICC. Sensitivity analyses were instead applied here, to evaluate the impact of uncertainty in the population ICC, consistent with the frequentist approach adopted.

Where the IPD are available, meta-regression analyses that incorporate treatment-by-covariate, therapist-by-covariate or indeed patient-by-covariate interactions may be of interest. Covariates may be available at the study-, therapist- or patient-levels. These explicitly account for unexplained variation at each level, and give appropriate weight to the studies, therapists and patients. They are therefore to be preferred over other analyses proposed in this context[363, 365, 367]. The ability to assess different predictors at each level also offers more flexibility, increasing the range of models that can be explored. While the increased sample sizes open up opportunities not usually present at a study-level, if the number of studies, or clusters per study, is small, computational problems may arise due to the presence of negative estimates, making additional assumptions necessary. The complexity of the model also needs balancing against the precision of its estimates. Analogous models for use when the IPD are unavailable would be useful only if cluster averages were reported alongside averages at the study-level. As this is unlikely to be typical in practice, interest in estimating interactions of this sort justifies collection of the IPD.

The focus here has been exclusively on meta-analyses of absolute mean differences in the context of a three-level model. It is important that issues arising specifically in the context of odds ratios, relative risks, hazard ratios, and standardised mean differences are considered too. Allowance for between-arm heteroscedasticity at multiple levels is less straightforward for these summary statistics. If an aggregate approach is adopted population averaged or marginal estimates are required, rather than cluster-specific or conditional ones[431]. Properties of the sampling distribution are also more complex[422]. Extensions that allow for further levels, such as centres or repeated observations over time, may also be important. These could be fit if the IPD were available.

6.6 Appendix: Programming Code for One-Step Models

STATA VERSION 11

The standard fixed-effects meta-analysis (model 6.21), for a dataset IPD_wide.dta with the variables study, treat and outcome can be fitted using the Stata code

```
use IPD_wide.dta, clear
xi: regress outcome i.study i.treat
```

The patient-level error can be allowed to differ by treatment, as in model (6.22), using

```
xi: xtmixed outcome i.study i.treat, resid(ind, by(treat))
```

The fixed-effects meta-analysis corresponding to the two-level heteroscedastic model, given in model (6.23), where t_id is the therapist identifier, can be fitted with

```
xi: xtmixed outcome i.study i.treat || t_id: treat, resid(ind, by(treat))
```

For partially-nested designs, this becomes

```
xi: xtmixed outcome i.study i.treat || t_id: treat, nocons resid(ind, by(treat))
```

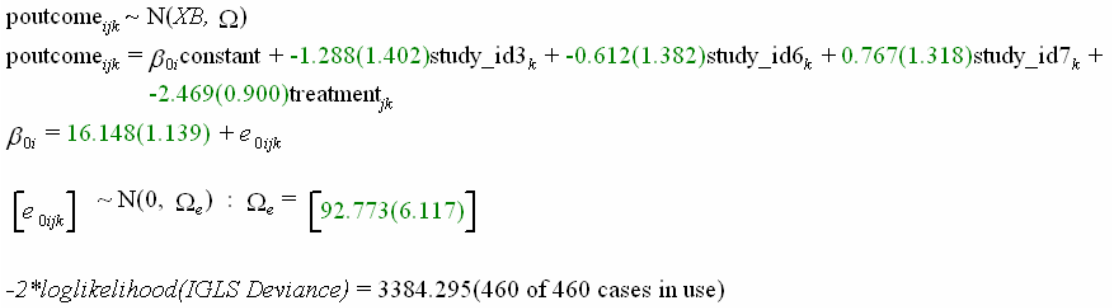
The random-effects meta-analysis, given in model (6.24), can be fitted using

```
xi: xtmixed outcome i.study i.treat || study: treat, nocons || t_id: treat, resid(ind, by(treat))
```

The meta-regression models in (6.25) to (6.29) cannot currently be fitted in Stata.

MLwiN VERSION 2.02

A dataset was imported starting with variables study_id, t_id, p_id identifying the study, cluster (i.e. counsellor or control patient), and patient, followed by indicator variables study_id2, study_id3, study_id6, study_id7 for Chilvers 2001, Friedli 1997, King 2000 and Simpson 2000, treatment for counselling, control for no counselling, poutcome for the BDI, and constant for a column of ones. The data were already sorted on study_id, t_id, and p_id and the data had been reduced to complete cases. Once in MLwiN, the RIGLS option under *Equations* was used, and the worksheet then saved. The *Equations* under *Model* was used to open an interactive window. The outcome was specified, and three levels. The standard fixed-effects meta-analysis (model 6.21), was fitted as follows:



$$\text{poutcome}_{ijk} \sim \text{N}(XB,\ \Omega)$$

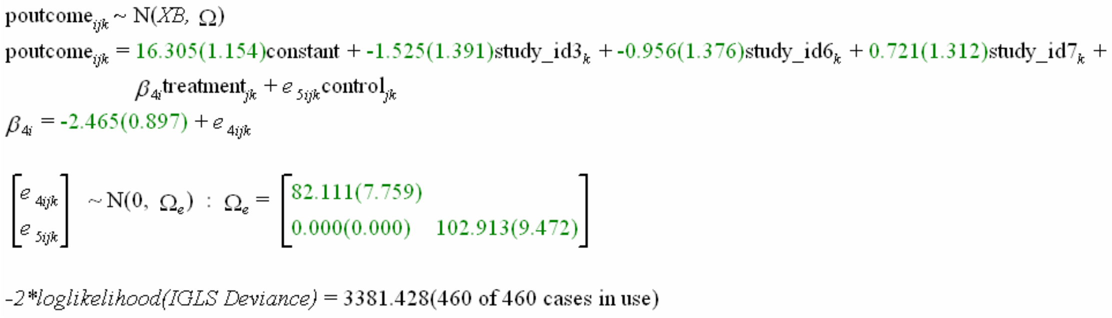
$$\text{poutcome}_{ijk} = \beta_{0i}\text{constant} + \text{-}1.288(1.402)\text{study\_id3}_k + \text{-}0.612(1.382)\text{study\_id6}_k + 0.767(1.318)\text{study\_id7}_k + \text{-}2.469(0.900)\text{treatment}_{jk}$$

$$\beta_{0i} = 16.148(1.139) + e_{0ijk}$$

$$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 92.773(6.117) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3384.295(460 of 460 cases in use)

The patient-level error was allowed to differ by treatment, as in model (6.22), using



$$\text{poutcome}_{ijk} \sim \text{N}(XB,\ \Omega)$$

$$\text{poutcome}_{ijk} = 16.305(1.154)\text{constant} + \text{-}1.525(1.391)\text{study\_id3}_k + \text{-}0.956(1.376)\text{study\_id6}_k + 0.721(1.312)\text{study\_id7}_k + \beta_{4i}\text{treatment}_{jk} + e_{5ijk}\text{control}_{jk}$$

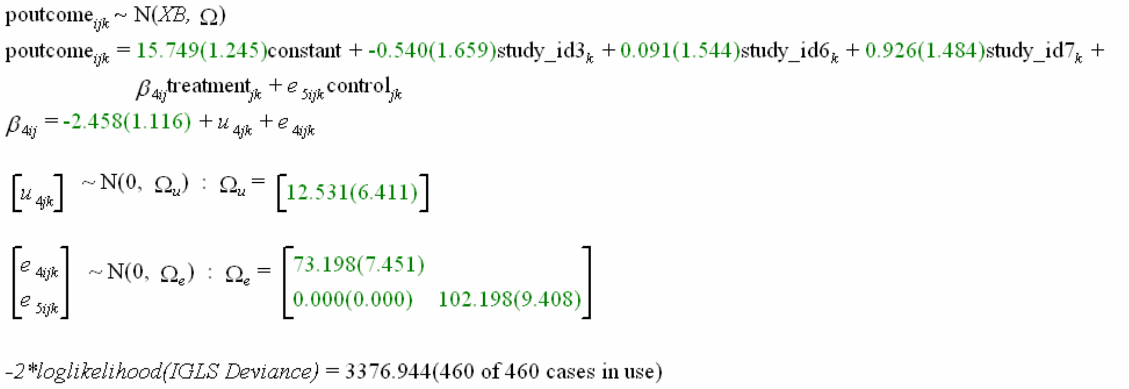
$$\beta_{4i} = \text{-}2.465(0.897) + e_{4ijk}$$

$$\begin{bmatrix} e_{4ijk} \\ e_{5ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 82.111(7.759) & \\ 0.000(0.000) & 102.913(9.472) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3381.428(460 of 460 cases in use)

where the constant is a fixed parameter only, while the control indicator variable is a level 1 term without a fixed parameter.

The fixed-effects meta-analysis corresponding to the two-level heteroscedastic model, given in model (6.23), was fitted with



$$\text{poutcome}_{ijk} \sim \text{N}(XB,\ \Omega)$$

$$\text{poutcome}_{ijk} = 15.749(1.245)\text{constant} + \text{-}0.540(1.659)\text{study\_id3}_k + 0.091(1.544)\text{study\_id6}_k + 0.926(1.484)\text{study\_id7}_k + \beta_{4ij}\text{treatment}_{jk} + e_{5ijk}\text{control}_{jk}$$

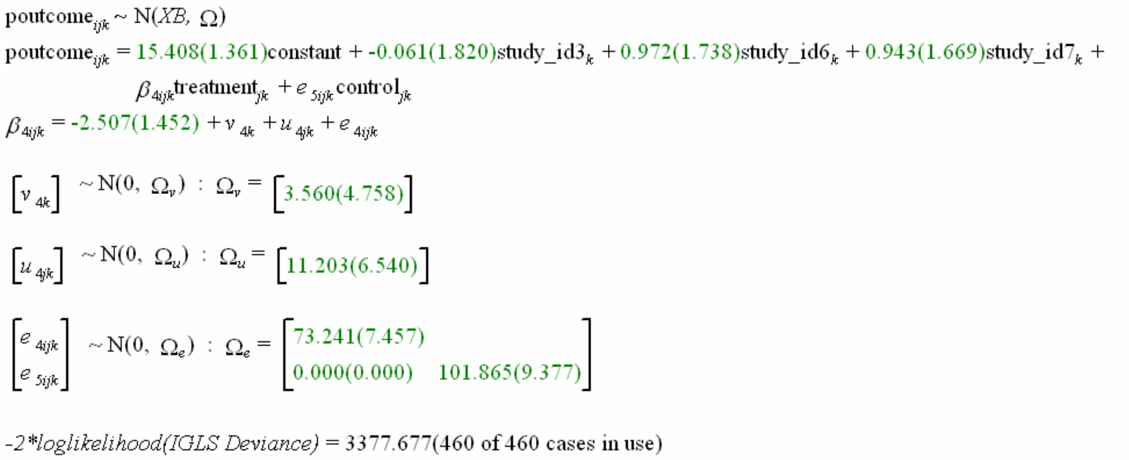
$$\beta_{4ij} = \text{-}2.458(1.116) + u_{4jk} + e_{4ijk}$$

$$\begin{bmatrix} u_{4jk} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 12.531(6.411) \end{bmatrix}$$

$$\begin{bmatrix} e_{4ijk} \\ e_{5ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 73.198(7.451) & \\ 0.000(0.000) & 102.198(9.408) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3376.944(460 of 460 cases in use)

while the random-effects meta-analysis, given in model (6.24), was fitted using



$$\text{poutcome}_{ijk} \sim \text{N}(XB,\ \Omega)$$

$$\text{poutcome}_{ijk} = 15.408(1.361)\text{constant} + \text{-}0.061(1.820)\text{study\_id3}_k + 0.972(1.738)\text{study\_id6}_k + 0.943(1.669)\text{study\_id7}_k + \beta_{4ijk}\text{treatment}_{jk} + e_{5ijk}\text{control}_{jk}$$

$$\beta_{4ijk} = \text{-}2.507(1.452) + v_{4k} + u_{4jk} + e_{4ijk}$$

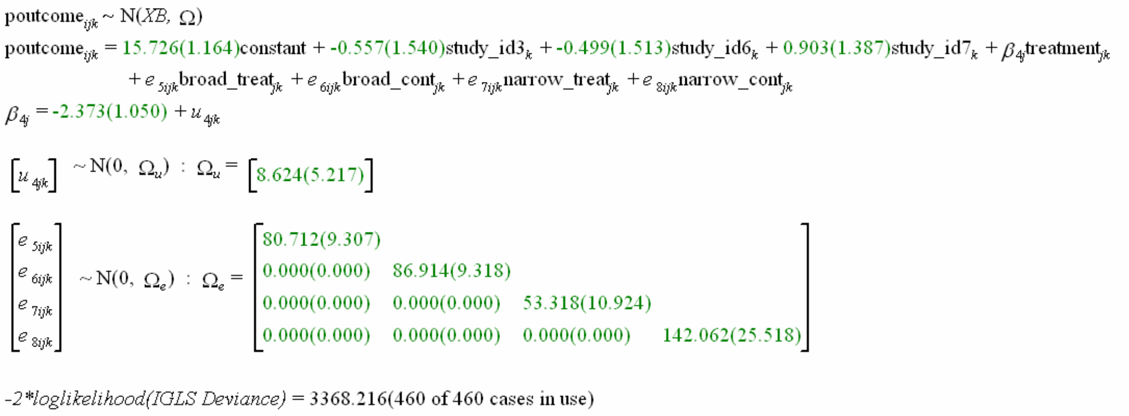
$$\begin{bmatrix} v_{4k} \end{bmatrix} \sim \text{N}(0,\ \Omega_v)\ :\ \Omega_v = \begin{bmatrix} 3.560(4.758) \end{bmatrix}$$

$$\begin{bmatrix} u_{4jk} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 11.203(6.540) \end{bmatrix}$$

$$\begin{bmatrix} e_{4ijk} \\ e_{5ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 73.241(7.457) & \\ 0.000(0.000) & 101.865(9.377) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3377.677(460 of 460 cases in use)

In order to fit the meta-regression models in Table 6.3, indicator variables broad_treat, broad_cont, narrow_treat, narrow_cont for the treatment-by-patient eligibility interaction and man_treat, noman_treat for the treatment-by-standardisation interaction were first added to the dataset. The fixed-effect meta-regression for the patient-level variance is given as:



$$\text{poutcome}_{ijk} \sim \text{N}(XB,\ \Omega)$$

$$\text{poutcome}_{ijk} = 15.726(1.164)\text{constant} + \text{-}0.557(1.540)\text{study\_id3}_k + \text{-}0.499(1.513)\text{study\_id6}_k + 0.903(1.387)\text{study\_id7}_k + \beta_{4j}\text{treatment}_{jk} + e_{5ijk}\text{broad\_treat}_{jk} + e_{6ijk}\text{broad\_cont}_{jk} + e_{7ijk}\text{narrow\_treat}_{jk} + e_{8ijk}\text{narrow\_cont}_{jk}$$

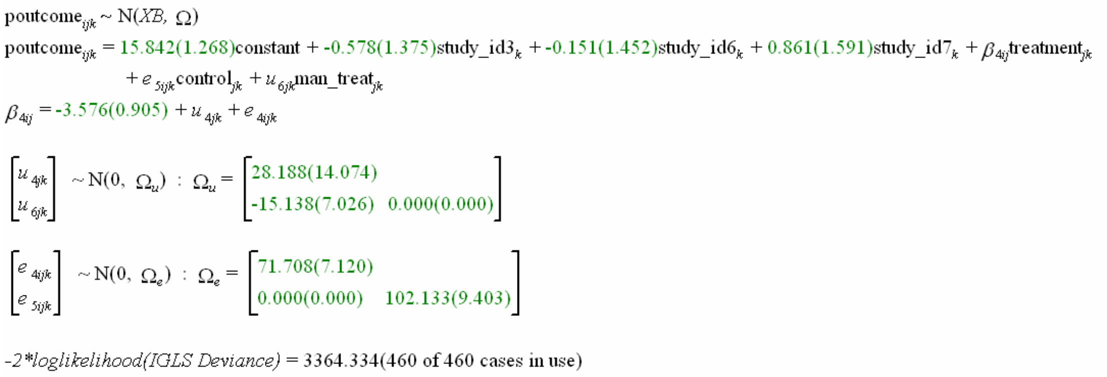
$$\beta_{4j} = \text{-}2.373(1.050) + u_{4jk}$$

$$\begin{bmatrix} u_{4jk} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 8.624(5.217) \end{bmatrix}$$

$$\begin{bmatrix} e_{5ijk} \\ e_{6ijk} \\ e_{7ijk} \\ e_{8ijk} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 80.712(9.307) & & & \\ 0.000(0.000) & 86.914(9.318) & & \\ 0.000(0.000) & 0.000(0.000) & 53.318(10.924) & \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 142.062(25.518) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3368.216(460 of 460 cases in use)

While the fixed-effect meta-regression for the counsellor-level variance is given as:



$$\text{poutcome}_{ijk} \sim \mathrm{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{poutcome}_{ijk} = {} & 15.842(1.268)\text{constant} + -0.578(1.375)\text{study\_id3}_k + -0.151(1.452)\text{study\_id6}_k + 0.861(1.591)\text{study\_id7}_k + \beta_{4ij}\text{treatment}_{jk} \\ & + e_{5ijk}\text{control}_{jk} + u_{6jk}\text{man\_treat}_{jk}\end{aligned}$$

$$\beta_{4ij} = -3.576(0.905) + u_{4jk} + e_{4ijk}$$

$$\begin{bmatrix} u_{4jk} \\ u_{6jk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 28.188(14.074) & \\ -15.138(7.026) & 0.000(0.000) \end{bmatrix}$$

$$\begin{bmatrix} e_{4ijk} \\ e_{5ijk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 71.708(7.120) & \\ 0.000(0.000) & 102.133(9.403) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3364.334(460 of 460 cases in use)

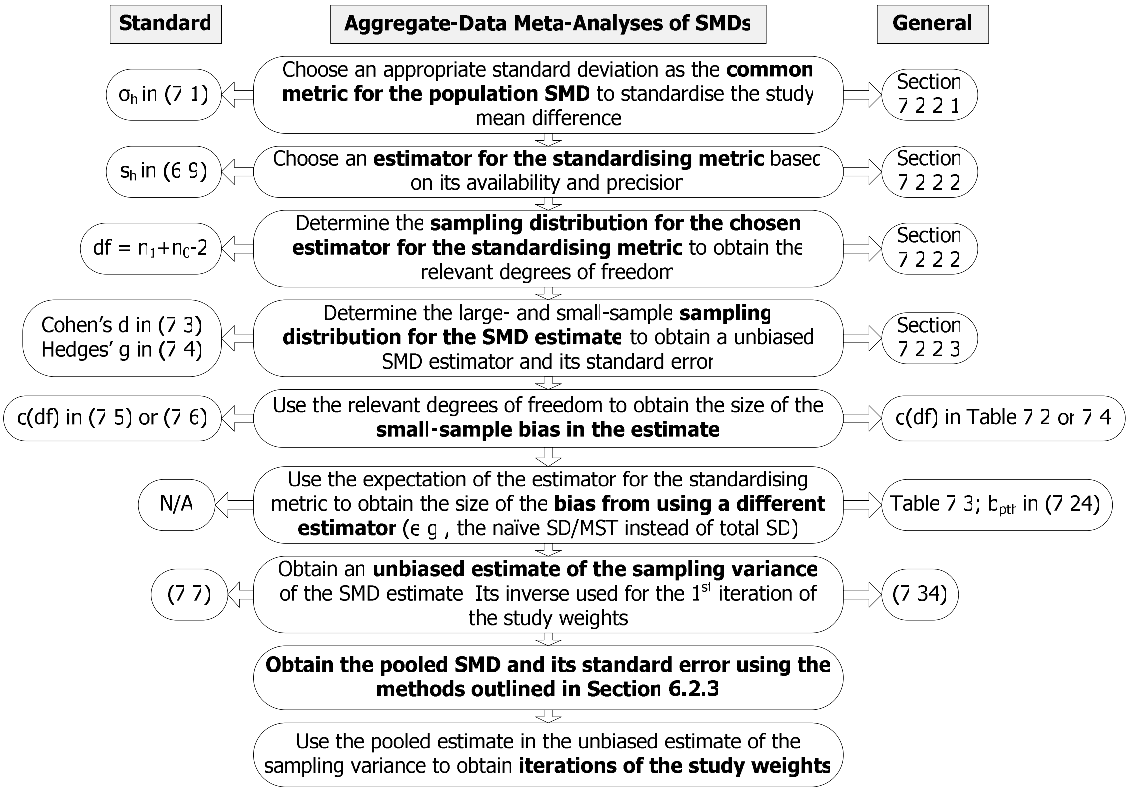
The parameterisation was altered to include a covariance term to avoid computational problems arising from a negative counsellor-level variance observed if counselling was standardised.

7 META-ANALYSIS OF STANDARDISED MEAN DIFFERENCES FROM NESTED THERAPIST DESIGNS

7.1 Introduction

If continuous outcomes are measured with different scales or standards across studies the relevant treatment effect may be the mean difference standardised to a common metric. Outcomes are then assumed to be linearly equitable across studies, regardless of the measurement tool used, and the treatment effect is interpreted as an absolute mean difference given in standard deviation units[432]. In the typical case, where study outcomes are independent within and across arms and their variance is homogeneous, the population standardised mean difference (SMD) is defined simply as the difference in means divided by the common standard deviation of the outcome. Where outcomes are clustered within arms, or outcome variances are heterogeneous across arms, there is no longer a single standard deviation, and the options available for scaling the mean difference can be numerous. In general, each alternative is associated with a different population SMD and therefore requires a different interpretation. Figure 7.1 gives an overview of the steps in a meta-analysis of SMDs based on the aggregate data.

Figure 7.1 Summary of Steps in Aggregate-Data Meta-Analyses of SMDs


Standard
Aggregate-Data Meta-Analyses of SMDs
General
σh in (7 1)
Choose an appropriate standard deviation as the common metric for the population SMD to standardise the study mean difference
Section 7 2 2 1
sh in (6 9)
Choose an estimator for the standardising metric based on its availability and precision
Section 7 2 2 2
df = n1+n0-2
Determine the sampling distribution for the chosen estimator for the standardising metric to obtain the relevant degrees of freedom
Section 7 2 2 2
Cohen's d in (7 3)
Hedges' g in (7 4)
Determine the large- and small-sample sampling distribution for the SMD estimate to obtain a unbiased SMD estimator and its standard error
Section 7 2 2 3
c(df) in (7 5) or (7 6)
Use the relevant degrees of freedom to obtain the size of the small-sample bias in the estimate
c(df) in Table 7 2 or 7 4
N/A
Use the expectation of the estimator for the standardising metric to obtain the size of the bias from using a different estimator (e g , the naïve SD/MST instead of total SD)
Table 7 3; bptr in (7 24)
(7 7)
Obtain an unbiased estimate of the sampling variance of the SMD estimate Its inverse used for the 1st iteration of the study weights
(7 34)
Obtain the pooled SMD and its standard error using the methods outlined in Section 6.2.3
Use the pooled estimate in the unbiased estimate of the sampling variance to obtain iterations of the study weights

In the context of the Behrens-Fisher problem[426], Glass[433, 434] argued that heterogeneity in treatments provided between studies complicates interpretation, and recommended the control arm standard deviation be used as the metric of choice if the comparator is no treatment. If the control content also varies from study to study, this advantage is lost. As an alternative, Huynh[41] suggested pooling the standard deviations, using the effect size proposed by Cohen[435] (p.44) where the sample size in each arm is assumed to be equal. While the resulting distribution is somewhat contrived and requires careful interpretation[436], this SMD has the advantage of reducing to the standard SMD when outcome variances are homogeneous. It also utilises all the available data, minimising the small-sample bias in the study estimates identified by Hedges[432, 437]. When sample sizes differ across arms, a more general pooled outcome variance could be used (see 6.8). Again the sample variances estimate different population variances in this case. A further option, if available, might be to use the associated baseline standard deviation. As a metric this is more usually recommended for standardised mean change scores[149-151, 438, 439], and assumes homoscedasticity across arms, but not necessarily across time, due to random allocation of treatments to patients[150, 151]. This may appeal particularly when the eligibility criteria are similar across the studies.

Where outcomes are clustered within arms but the outcome variances are assumed to be homogeneous across arms, a situation that is plausible in cluster-randomised trials, the total outcome variance is split into within- and between-cluster components. White and Thomas[42] and Hedges[43] have suggested three further population SMDs based on these standard deviations, respectively. When the ICC is known and between zero and one, they can be easily converted. This facilitates comparability in the definition of the SMD across studies. If there is only one cluster per arm, the between-cluster variance will not be defined. Likewise the within-cluster variance may be unavailable if analyses are reported at the cluster-level. As a result, assumptions could be made regarding the ICC in such studies, so that the SMD might be reported and interpreted in units of the total standard deviation. While the choice of metric depends on the inference that is of interest to the meta-analyst[43], SMDs based on the total and within standard deviations reduce to the usual SMD when outcomes are independent. If clustering was ignored in published analyses, estimates of the total, within and between standard deviations are unlikely to be readily available. White and Thomas[42] and Hedges[43] therefore suggested the usual or naïve standard deviation as a proxy for the total standard deviation when estimating the SMD based on the latter, allowing for an additional bias that arises in doing so.

Use of the individual-patient-data in meta-analyses of SMDs appears to be limited, but see[440-442] for examples. Goldstein *et al*[44] described a one-step approach, suggesting the level-1 or within-cluster standard deviation as the common metric. This was illustrated using studies of class sizes, in which students were nested within classes, schools and studies and small versus large class sizes represented the treatment arms. Inclusion of a further level in the meta-analysis between the classes and studies makes Goldstein *et al*s[44] approach very relevant. However the schools were crossed with the treatment arms in their example, and while they alluded to more complex models which allow for between-arm or study heteroscedasticity, they did not consider nested study designs; the rationale for, or implications of, the choice of metric; allowance for imprecision in a standardising standard deviation; or the relationship between the means by which the data are standardised and the choice of model for the meta-analysis.

Nested therapist designs are characterised by between-arm heteroscedasticity at both the therapist and patient levels[84, 286]. Methods are therefore needed that relax *both* the independence and common variance assumptions for the studies, allowing the sample sizes to differ across arms. The aim of this chapter was accordingly to adapt, illustrate and compare aggregate and one-step methods for pooling SMDs from nested therapist designs. The mixture of designs employed in a psychotherapy context (see Chapter 3), implies a general metric, with an equivalent interpretation across study designs, would be attractive. The use of the pooled total, naïve or within-cluster standard deviation in an aggregate setting acts as a natural extension of Huynh[41], White and Thomas[42] and Hedges[43]. As the within-cluster standard deviation is not defined for the control arm of studies with partially nested designs, methods relating to the pooled total SMD have a greater potential in this regard. However, since the total and naïve standard deviations are affected by within-study clustering, their SMD estimate is too. This complicates the meta-analysis, and makes its interpretation more difficult, particularly if the underlying ICCs are believed to vary from study to study. Given that Glass' SMD avoids the impact of clustering on the study estimate in the context of partially nested designs, it has the potential to avoid some of the additional complexities found for the pooled total SMD. Its use is however restricted to the special case where studies have partially nested or independent designs.

The chapter is structured as follows. Section 7.2 sets out the aggregate-data methods, initially outlining standard fixed- and random-effects meta-analyses and the distinction between Cohen's d and Hedges' g, and then proposing a more general approach that simultaneously allows for within-study clustering and between-arm heteroscedasticity. Section 7.3 sets out the one-step methods, initially describing the means by which the outcomes are transformed, and then outlining the multilevel meta-analysis models that correspond to different choices of standardising metric. These are illustrated in Section 7.4 using the counselling in primary care example introduced in Chapter 4.

7.2 Aggregate-Data Methods

7.2.1 Standard Fixed- and Random-Effects Meta-Analysis Models

Where the independence and common variance assumptions hold for all the studies to be combined, the population SMD is defined as[145]

$$\theta_{SMD,h} = \frac{\mu_{h1} - \mu_{h0}}{\sigma_h}, \quad h = 1, \ldots, H \qquad (7.1)$$

with the difference in the population means of the intervention and control arms of the h^{th} study given by $\mu_{h1} - \mu_{h0}$, and the common metric by the standard deviation of the outcome, σ_h. The population metric could be estimated by s_{h1}, s_{h0} or by s_h, where s denotes the sample estimate. However, the pooled standard deviation, s_h, maximises the degrees of freedom available. As $s_h^2 = MSE_h$ (see 6.9), it follows that the sampling distribution of s_h^2 is exactly proportional to a chi-square with $n_{h1} + n_{h0} - 2$ degrees of freedom. Where the study sample sizes are large, so that all of the $df_{s_h^2}$ are also large, the study estimate of $\theta_{SMD,h}$ is given simply by Cohen's d,

$$\hat{\theta}_{Cohen's\ d,h} = \frac{\bar{y}_{h1} - \bar{y}_{h0}}{s_h} \qquad (7.2)$$

where $\bar{y}_{h1} - \bar{y}_{h0}$ is the difference in the sample means, and the sampling distribution is given asymptotically by[148, 432, 437]

$$\hat{\theta}_{Cohen's\ d,h} \sim N\left(\frac{\mu_{h1}-\mu_{h0}}{\sigma_h}, \left(\frac{1}{n_{h1}}+\frac{1}{n_{h0}}\right)+\frac{\theta^2_{SMD,h}}{2\left(n_{h1}+n_{h0}\right)}\right) \qquad (7.3)$$

If the study sample sizes are small, and particularly if $df_{\{s_h^2\}} \leq 10$, Hedges[432, 437] showed that Cohen's d is biased for $\theta_{SMD,h}$, and derived an alternative estimator to correct for this. He gave Hedges' g as

$$\hat{\theta}_{Hedges'\ g,h} = c(df)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_h}\right) \sim N\left(\frac{\mu_{h1}-\mu_{h0}}{\sigma_h}, \frac{c(df)^2 df\left(1+\tilde{n}_h\theta^2_{SMD,h}\right)}{(df-2)\tilde{n}_h} - \theta^2_{SMD,h}\right) \qquad (7.4)$$

where $\tilde{n}_h = \dfrac{n_{h1}n_{h0}}{n_{h1}+n_{h0}} = \left(\dfrac{1}{n_{h1}}+\dfrac{1}{n_{h0}}\right)^{-1}$ and $c(df) = \Gamma\left(\dfrac{df}{2}\right)\Big/\left(\sqrt{\dfrac{df}{2}}\,\Gamma\left(\dfrac{df-1}{2}\right)\right)$ (7.5)

An approximation for the small-sample correction $c(df)$ was given by Hedges[432, 437] as

$$c(df) \approx 1-\frac{3}{4df-1} \qquad (7.6)$$

Hedges' g is unbiased, regardless of the degrees of freedom available to estimate s_h^2, so is preferred over Cohen's d. Since $c(\infty)=1$, Hedges' g converges to Cohen's d as study sample sizes increase, but it is uniformly smaller than Cohen's d otherwise[432, 437]. Thus, the difference between these estimators is only important when the degrees of freedom available for estimating the standardising metric are very small, in one or more of the studies.

It is evident from (7.3) and (7.4) that the sampling variance of Cohen's d and Hedges' g is a function of the squared parameter, $\theta^2_{SMD,h}$. Hedges[432, 437] suggested substituting the squared sample estimate when estimating the standard error, White and Thomas[42] (p.150) show that this introduces bias. This is because the expectation of the squared estimate is equal to the squared parameter *plus* the variance of the estimate. That is, $E\left(\hat{\theta}^2\right) = \theta^2 + \sigma^2_{\{\hat{\theta}\}}$. They proposed a refined estimator for the exact variance, given by

$$\hat{\sigma}^2_{\{\hat{\theta}_{Hedges'\ g,h}\}} = \left(\frac{1}{n_{h1}}+\frac{1}{n_{h0}}\right) + \hat{\theta}^2_{Hedges'\ g,h}\left(1-\frac{df-2}{c(df)^2 df}\right) \qquad (7.7)$$

where the first term relates to the variance of the numerator of the study estimate and the second to the variance of its denominator. This estimator was originally derived by Hedges[443] (p.391).

It is clear that the standard error of an SMD increases as a function of its expectation, mainly if the degrees of freedom available for estimating its denominator are low. This is problematic if a common metric is assumed across studies, as in standard fixed- and random-effects meta-analysis models (Models 6.1 and 6.2). It has a similar effect here as it did for ICCs (Section 5.3.1), i.e. the pooled estimate may be unduly affected by a single study with a small SMD. Hedges[437] argued that when the denominator degrees of freedom are large this can be ignored. If they are not Hedges[432] advised modifying the estimate of the sampling variance used for study weights, replacing the population value by the pooled estimate and iterating, particularly when the SMDs vary across the studies. As with other aggregate meta-analyses, it is common to assume the sampling variance is known when estimating the study weights[437]. A robust sandwich estimator proposed by Sidik and Jonkman[40] to protect against imprecision in the study weights is also applicable in this context (see Section 6.2.3).

7.2.2 General Fixed- and Random-Effects Meta-Analysis Models

7.2.2.1 Choice of Metric for the Population Standardised Mean Difference

The general population SMD is defined as

$$\theta_{SMD,h} = \frac{\mu_{h1} - \mu_{h0}}{\sigma_{den,h}} \qquad (7.8)$$

where the form of $\sigma_{den,h}$ depends on the choice of metric. So, for example, $\sigma_{den,h} = \sigma_{h0}$ for Glass's[433, 434] SMD. It is equal to $\sqrt{\left(\sigma_{h1}^2 + \sigma_{h0}^2\right)/2}$ for Huynh's[41] SMD, and to σ_{th}, σ_{wh} or σ_{bh} for SMDs using the total, within- or between-cluster standard deviations[42, 43].

The most general standardising metric for studies with nested therapist designs is the pooled total standard deviation. The SMD based on this is given by

$$\theta_{pth} = \frac{\mu_{h1} - \mu_{h0}}{\sqrt{\frac{(n_{h1}-1)\sigma_{th1}^2 + (n_{h0}-1)\sigma_{th0}^2}{n_{h1}+n_{h0}-2}}} \qquad (7.9)$$

where the cluster sizes are assumed to be equal within the arms, and n denotes the number of patients. This simplifies to

$$\theta_{pth|\rho_{h0}=0} = \frac{\mu_{h1} - \mu_{h0}}{\sqrt{\frac{(n_{h1}-1)\sigma_{th1}^2 + (n_{h0}-1)\sigma_{h0}^2}{n_{h1}+n_{h0}-2}}} \qquad (7.10)$$

for partially nested designs, and to

$$\theta_{ph} = \frac{\mu_{h1} - \mu_{h0}}{\sqrt{\frac{(n_{h1}-1)\sigma_{h1}^2 + (n_{h0}-1)\sigma_{h0}^2}{n_{h1}+n_{h0}-2}}} \qquad (7.11)$$

for the Behrens-Fisher problem. When the sample sizes are equal across arms, (7.11) simplifies further to Huynh's[41] SMD. If the sample sizes and the outcome variances are equal across the arms, (7.9) reduces to the SMD based on σ_{th}, described by White and Thomas[42] and Hedges[43]. As such, the standard case, the Behrens-Fisher problem, and other two-level nested designs can all be viewed as special cases. While meta-analyses of mixed designs are likely to lead to systematic variation in the estimates across study designs, the use of a general metric does ensure their interpretation is comparable.

7.2.2.2 Sampling Distributions for the Standardising Standard Deviations

The sample estimate of the pooled total variance is given by

$$s_{pth}^2 = \frac{(n_{h1}-1)s_{th1}^2 + (n_{h0}-1)s_{th0}^2}{n_{h1}+n_{h0}-2} \qquad (7.12)$$

As $s_{thi}^2 = \frac{SSB_{hi}}{m_{hi}(k_{hi}-1)} + \frac{SSW_{hi}}{k_{hi}m_{hi}} = \left(\frac{1}{m_{hi}}\right)MSB_{hi} + \left(\frac{m_{hi}-1}{m_{hi}}\right)MSW_{hi}$

It follows that

$$s_{pth}^{2}=\sum_{i=0}^{1}\left(\left(\frac{n_{hi}-1}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSB_{hi}+\left(\frac{\left(n_{hi}-1\right)\left(m_{hi}-1\right)}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSW_{hi}\right)$$

Using a Satterthwaite approximation[108], the distribution of s_{pth}^{2} can be approximated to a chi-square with degrees of freedom given by

$$df_{\left\{s_{pth}^{2}\right\}}=\frac{\left(\sum_{i=0}^{1}\left(\left(\frac{n_{hi}-1}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSB_{hi}+\left(\frac{\left(n_{hi}-1\right)\left(m_{hi}-1\right)}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSW_{hi}\right)\right)^{2}}{\sum_{i=0}^{1}\left(\frac{\left(\left(\frac{n_{hi}-1}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSB_{hi}\right)^{2}}{\left(k_{hi}-1\right)}+\frac{\left(\left(\frac{\left(n_{hi}-1\right)\left(m_{hi}-1\right)}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)MSW_{hi}\right)^{2}}{k_{hi}\left(m_{hi}-1\right)}\right)}$$

Since $MSB_{hi}=m_{hi}s_{bhi}^{2}+s_{whi}^{2}$, $s_{bhi}^{2}=\frac{\rho_{hi}s_{whi}^{2}}{(1-\rho_{hi})}$ and $MSW_{hi}=s_{whi}^{2}$

It follows that $MSB_{hi}=m_{hi}\left(\frac{\rho_{hi}s_{whi}^{2}}{(1-\rho_{hi})}\right)+s_{whi}^{2}$ and that

$$df_{\left\{s_{pth}^{2}\right\}}=\frac{\left(\sum_{i=0}^{1}\left(\left(\frac{n_{hi}-1}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)\left(m_{hi}\left(\frac{\rho_{hi}s_{whi}^{2}}{(1-\rho_{hi})}\right)+s_{whi}^{2}\right)+\left(\frac{\left(n_{hi}-1\right)\left(m_{hi}-1\right)}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)s_{whi}^{2}\right)\right)^{2}}{\sum_{i=0}^{1}\left(\frac{\left(\left(\frac{n_{hi}-1}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)\left(m_{hi}\left(\frac{\rho_{hi}s_{whi}^{2}}{(1-\rho_{hi})}\right)+s_{whi}^{2}\right)\right)^{2}}{\left(k_{hi}-1\right)}+\frac{\left(\left(\frac{\left(n_{hi}-1\right)\left(m_{hi}-1\right)}{m_{hi}\left(n_{h1}+n_{h0}-2\right)}\right)s_{whi}^{2}\right)^{2}}{k_{hi}\left(m_{hi}-1\right)}\right)}$$

This simplifies to give

$$df_{\left\{s_{pth}^{2}\right\}}=\frac{\left(\left(n_{h1}-1\right)\left(\rho_{h2}-1\right)s_{wh1}^{2}+\left(n_{h2}-1\right)\left(\rho_{h1}-1\right)s_{wh2}^{2}\right)^{2}}{\left(\rho_{h1}-1\right)^{2}\left(\rho_{h2}-1\right)^{2}\left(\sum_{i=0}^{1}\left(\frac{\left(n_{hi}-1\right)^{2}}{m_{hi}^{2}}\left(\frac{m_{hi}-1}{k_{hi}}+\frac{\left(1+\left(m_{hi}-1\right)\rho_{hi}\right)^{2}}{\left(k_{hi}-1\right)\left(\rho_{hi}-1\right)^{2}}\right)s_{whi}^{4}\right)\right)} \qquad (7.13)$$

This, in turn, simplifies to give

$$df_{\left\{s_{pth|\rho_{h0}=0}^{2}\right\}}=\frac{\left(\left(1-n_{h1}\right)s_{wh1}^{2}+\left(n_{h0}-1\right)\left(\rho_{h1}-1\right)s_{h0}^{2}\right)^{2}}{\left(\rho_{h1}-1\right)^{2}\left(\left(n_{h0}-1\right)^{2}s_{h0}^{4}+\left(\frac{\left(n_{h1}-1\right)^{2}}{m_{h1}^{2}}\left(\frac{m_{h1}-1}{k_{h1}}+\frac{\left(1+\left(m_{h1}-1\right)\rho_{h1}\right)^{2}}{\left(k_{h1}-1\right)\left(\rho_{h1}-1\right)^{2}}\right)s_{wh1}^{4}\right)\right)} \qquad (7.14)$$

where $\rho_{h0}=0$, $k_{h0}=n_{h0}$ and $m_{h0}=1$ for a partially nested design.

Tables 7.1 and 7.2 summarise the sample estimates for the pooled total variance and their sampling distributions, respectively, under the other more restrictive scenarios. In each case the general formula for the degrees of freedom (7.13) simplifies to the more specific one. Under the most restrictive assumptions, (7.13) reduces to $n_{h1} + n_{h0} - 2$. If the common variance assumption is relaxed, (7.13) reduces to the degrees of freedom Huynh[41] (p.21) gave for the Behrens-Fisher problem. If the independence assumption is instead relaxed, (7.13) reduces to the degrees of freedom under a random-intercept model. This corrects typographical errors in White and Thomas[42] (p.151) and Hedges[43] (p.364). The other degrees of freedom relate to (7.9), (7.10) and (7.11), respectively.

The general form of the pooled naïve variance is

$$s_{ph}^2 = \frac{SSW_{h1} + SSB_{h1} + SSW_{h0} + SSB_{h0}}{k_{h1}m_{h1} + k_{h0}m_{h0} - 2} \qquad (7.15)$$

assuming the cluster sizes are equal within arms. Its expectation under a two-level heteroscedastic model is

$$\begin{aligned}
E\left[s_{ph}^2\right] &= E\left[MST_h\right] = E\left[\frac{SST_h}{N_h - 2}\right] = E\left[\frac{SSB_{h1} + SSW_{h1} + SSB_{h0} + SSW_{h0}}{n_{h1} + n_{h0} - 2}\right] \\
&= \frac{(k_{h1}-1)\left(\sigma_{wh1}^2 + m_{h1}\sigma_{bh1}^2\right) + k_{h1}(m_{h1}-1)\sigma_{wh1}^2 + (k_{h0}-1)\left(\sigma_{wh0}^2 + m_{h0}\sigma_{bh0}^2\right) + k_{h0}(m_{h0}-1)\sigma_{wh0}^2}{n_{h1} + n_{h0} - 2} \\
&= \frac{m_{h1}(k_{h1}-1)\sigma_{bh1}^2 + (k_{h1}m_{h1}-1)\sigma_{wh1}^2 + m_{h0}(k_{h0}-1)\sigma_{bh0}^2 + (k_{h0}m_{h0}-1)\sigma_{wh0}^2}{n_{h1} + n_{h0} - 2} \\
&= \frac{(n_{h1}-1)\sigma_{th1}^2 + (n_{h0}-1)\sigma_{th0}^2 - (m_{h1}-1)\rho_{h1}\sigma_{th1}^2 - (m_{h0}-1)\rho_{h0}\sigma_{th0}^2}{n_{h1} + n_{h0} - 2} \\
&= \sigma_{pth}^2\left(1 - \frac{(m_{h1}-1)\rho_{h1}\sigma_{th1}^2 + (m_{h0}-1)\rho_{h0}\sigma_{th0}^2}{(n_{h1}-1)\sigma_{th1}^2 + (n_{h0}-1)\sigma_{th0}^2}\right) \qquad (7.16)
\end{aligned}$$

The pooled naïve standard deviation and its expectation are summarised in Table 7.3 for other more restrictive situations, including that of partial nesting. The expectations of the total and naïve standard deviations are identical if the independence assumption holds. If clustering is present, the naïve variance underestimates the total variance by a factor linked to the design effect, denoted b by Hedges[43].

As $MSW_{hi} = \dfrac{SSW_{hi}}{k_{hi}(m_{hi}-1)}$ and $MSB_{hi} = \dfrac{SSB_{hi}}{k_{hi}-1}$, it follows that

Table 7.1 Family of Total Standard Deviations under Various Model Assumptions

Assumptions			Standardising Metric	
Outcome Variances	**Clustering**	**Sample Sizes**	$s_{den,h}$	**as Sums of Squares**
$\sigma_{h1}^2 = \sigma_{h0}^2 = \sigma_h^2$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	$s_h = \sqrt{\frac{(n_{h1}-1)s_{h1}^2 + (n_{h0}-1)s_{h0}^2}{n_{h1}+n_{h0}-2}}$	$\sqrt{\frac{SSE_h}{n_{h1}+n_{h0}-2}}$
$\sigma_{h1}^2 \neq \sigma_{h0}^2$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} = n_{h0}$	$s_{Huynh,h} = \sqrt{\frac{s_{h1}^2 + s_{h0}^2}{2}}$	$\sqrt{\frac{SSE_{h1} + SSE_{h0}}{n_h - 1}}$
$\sigma_{h1}^2 \neq \sigma_{h0}^2$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	$s_{ph} = \sqrt{\frac{(n_{h1}-1)s_{h1}^2 + (n_{h0}-1)s_{h0}^2}{n_{h1}+n_{h0}-2}}$	$\sqrt{\frac{SSE_{h1} + SSE_{h1}}{n_{h1}+n_{h0}-2}}$
$\sigma_{bh1}^2 = \sigma_{bh0}^2 = \sigma_{bh}^2$ $\sigma_{wh1}^2 = \sigma_{wh0}^2 = \sigma_{wh}^2$ $\sigma_{th1}^2 = \sigma_{th0}^2 = \sigma_{th}^2$	$\rho_{h1} = \rho_{h0} = \rho_h$	$k_{h1} = k_{h0} = k_h$ $m_{h1} = m_{h0} = m_h$ so $n_{h1} = n_{h0}$	$s_{th} = \sqrt{s_{wh}^2 + s_{bh}^2}$	$\sqrt{\frac{SSW_h}{2k_h m_h} + \frac{SSB_h}{2m_h(k_h-1)}}$
$\sigma_{bh0}^2 = 0$ $\sigma_{wh1}^2 \neq \sigma_{h0}^2$ $\sigma_{th1}^2 \neq \sigma_{h0}^2$	$\rho_{h0} = 0$ $\rho_{h1} \neq 0$	$k_{h0} = n_{h0}$ $m_{h0} = 1$ $n_{h1} \neq n_{h0}$	$s_{pth\|\rho_{h0}=0} = \sqrt{\frac{(n_{h1}-1)s_{th1}^2 + (n_{h0}-1)s_{h0}^2}{n_{h1}+n_{h0}-2}}$	$\sqrt{\frac{(n_{h1}-1)\left(\frac{SSW_{h1}}{k_{h1}m_{h1}} + \frac{SSB_{h1}}{m_{h1}(k_{h1}-1)}\right) + SSE_{h0}}{n_{h1}+n_{h0}-2}}$
$\sigma_{bh1}^2 \neq \sigma_{bh0}^2$ $\sigma_{wh1}^2 \neq \sigma_{wh0}^2$ $\sigma_{th1}^2 \neq \sigma_{th0}^2$	$\rho_{h1} \neq \rho_{h0}$	$k_{h1} \neq k_{h0}$ $m_{h1} \neq m_{h0}$ $n_{h1} \neq n_{h0}$	$s_{pth} = \sqrt{\frac{(n_{h1}-1)s_{th1}^2 + (n_{h0}-1)s_{th0}^2}{n_{h1}+n_{h0}-2}}$	$\sqrt{\frac{\sum_{i=0}^{1}(n_{hi}-1)\left(\frac{SSW_{hi}}{k_{hi}m_{hi}} + \frac{SSB_{hi}}{m_{hi}(k_{hi}-1)}\right)}{n_{h1}+n_{h0}-2}}$

Note σ = the population outcome variance, s = its sample estimate, ρ = the population intraclass correlation, k = the number of clusters, m = the cluster size, n = the sample size, h = study, i = arm, b = between, t = total, w = within, e = error, SS = sums of squares

Table 7.2 Sampling Distributions for the Family of Total Variances

Assumptions			Degrees of Freedom	
Outcome Variances	Clustering	Sample Sizes	$s^2_{den,h}$	$df_{s^2_{den,h}}$
$\sigma^2_{h1} = \sigma^2_{h0} = \sigma^2_h$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	s^2_h	$n_{h1} + n_{h0} - 2$
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} = n_{h0}$	$s^2_{Huynh,h}$	$\dfrac{(n_{h1}-1)(n_{h0}-1)(\sigma^2_{h1}+\sigma^2_{h0})^2}{(n_{h0}-1)\sigma^4_{h1}+(n_{h1}-1)\sigma^4_{h0}}$
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	s^2_{ph}	$\dfrac{((n_{h1}-1)\sigma^2_{h1}+(n_{h0}-1)\sigma^2_{h0})^2}{(n_{h1}-1)\sigma^4_{h1}+(n_{h0}-1)\sigma^4_{h0}}$
$\sigma^2_{bh1} = \sigma^2_{bh0} = \sigma^2_{bh}$ $\sigma^2_{wh1} = \sigma^2_{wh0} = \sigma^2_{wh}$ $\sigma^2_{th1} = \sigma^2_{th0} = \sigma^2_{th}$	$\rho_{h1} = \rho_{h0} = \rho_h$	$k_{h1} = k_{h0} = k_h$ $m_{h1} = m_{h0} = m_h$ so $n_{h1} = n_{h0}$	s^2_{th}	$\dfrac{2k_h m^2_h (k_h-1)}{(k_h-1)(1-\rho_h)^2(m_h-1)+k_h(1+\rho_h(m_h-1))^2}$
$\sigma^2_{bh0} = 0$ $\sigma^2_{wh1} \neq \sigma^2_{h0}$ $\sigma^2_{th1} \neq \sigma^2_{h0}$	$\rho_{h0} = 0$ $\rho_{h1} \neq 0$	$k_{h0} = n_{h0}$ $m_{h0} = 1$ $n_{h1} \neq n_{h0}$	$s^2_{pth\mid\rho_{h0}=0}$	$\dfrac{((1-n_{h1})\sigma^2_{wh1}+(n_{h0}-1)(\rho_{h1}-1)\sigma^2_{h0})^2}{(\rho_{h1}-1)^2\left(\left(\frac{(n_{h1}-1)^2}{m^2_{h1}}\left(\frac{m_{h1}-1}{k_{h1}}+\frac{(1+(m_{h1}-1)\rho_{h1})^2}{(k_{h1}-1)(\rho_{h1}-1)^2}\right)\sigma^4_{wh1}\right)+(n_{h0}-1)^2\sigma^4_{h0}\right)}$
$\sigma^2_{bh1} \neq \sigma^2_{bh0}$ $\sigma^2_{wh1} \neq \sigma^2_{wh0}$ $\sigma^2_{th1} \neq \sigma^2_{th0}$	$\rho_{h1} \neq \rho_{h0}$	$k_{h1} \neq k_{h0}$ $m_{h1} \neq m_{h0}$ $n_{h1} \neq n_{h0}$	s^2_{pth}	$\dfrac{((n_{h1}-1)(\rho_{h2}-1)\sigma^2_{wh1}+(n_{h2}-1)(\rho_{h1}-1)\sigma^2_{wh2})^2}{(\rho_{h1}-1)^2(\rho_{h2}-1)^2\left(\sum_{i=0}^{1}\left(\frac{(n_{hi}-1)^2}{m^2_{hi}}\left(\frac{m_{hi}-1}{k_{hi}}+\frac{(1+(m_{hi}-1)\rho_{hi})^2}{(k_{hi}-1)(\rho_{hi}-1)^2}\right)\sigma^4_{whi}\right)\right)}$

Note σ = the population outcome variance, s = its sample estimate, ρ = the population intraclass correlation, k = the number of clusters, m = the cluster size, n = the sample size, h = study, i = arm, b = between, t = total, w = within

$$s_{ph}^2 = \sum_{i=0}^{1}\left(\left(\frac{k_{hi}-1}{n_{h1}+n_{h0}-2}\right)MSB_{hi} + \left(\frac{k_{hi}\left(m_{hi}-1\right)}{n_{h1}+n_{h0}-2}\right)MSW_{hi}\right) \qquad (7.17)$$

The pooled naïve variance is thus distributed proportionately to a chi-square according to Satterthwaite[108], with approximate degrees of freedom given by

$$df_{\left\{s_{ph}^2\right\}} = \frac{\left(\left(\left(n_{h1}-1\right)-\left(m_{h1}-1\right)\rho_{h1}\right)\left(\rho_{h2}-1\right)\sigma_{wh1}^2 + \left(\left(n_{h2}-1\right)-\left(m_{h2}-1\right)\rho_{h2}\right)\left(\rho_{h1}-1\right)\sigma_{wh2}^2\right)^2}{\left(\rho_{h1}-1\right)^2\left(\rho_{h2}-1\right)^2\left(\sum_{i=0}^{1}\sigma_{whi}^4\left(k_{hi}\left(m_{hi}-1\right)+\frac{\left(k_{hi}-1\right)\left(1+\left(m_{hi}-1\right)\rho_{hi}\right)^2}{\left(\rho_{hi}-1\right)^2}\right)\right)} \qquad (7.18)$$

Rearranging (5.26) gives

$$\sigma_{whi}^2 = \frac{\left(1-\rho_{hi}\right)\sigma_{hi}^2}{\left(1-\frac{\left(m_{hi}-1\right)\rho_{hi}}{n_{hi}-1}\right)} \qquad (7.19)$$

and substituting (7.19) into (7.18) gives

$$df_{\left\{s_{ph}^2\right\}} = \frac{\left(\sum_{i=0}^{1}\left(n_{hi}-1\right)\sigma_{hi}^2\right)^2}{\sum_{i=0}^{1}\left(\frac{\left(n_{hi}-1\right)^2\sigma_{hi}^4\left(n_{hi}\left(1+\left(m_{hi}-1\right)\rho_{hi}^2\right)-\left(1+\left(m_{hi}-1\right)\rho_{hi}\right)^2\right)}{\left(\left(n_{hi}-1\right)-\left(m_{hi}-1\right)\rho_{hi}\right)^2}\right)} \qquad (7.20)$$

which, in turn, reduces to

$$df_{\left\{s_{ph|\rho_{h0}=0}^2\right\}} = \frac{\left(\left(n_{h1}-1\right)\sigma_{h1}^2+\left(n_{h0}-1\right)\sigma_{h0}^2\right)^2}{\frac{\left(n_{h1}-1\right)^2\sigma_{h1}^4\left(n_{h1}\left(1+\left(m_{h1}-1\right)\rho_{h1}^2\right)-\left(1+\left(m_{h1}-1\right)\rho_{h1}\right)^2\right)}{\left(\left(n_{h1}-1\right)-\left(m_{h1}-1\right)\rho_{h1}\right)^2}+\left(n_{h0}-1\right)\sigma_{h0}^4} \qquad (7.21)$$

where $\rho_{h0}=0$, $k_{h0}=n_{h0}$ and $m_{h0}=1$ for a partially nested design.

Table 7.4 summarises the sampling distributions for the more restrictive scenarios. As expected, under independence, the degrees of freedom are the same for the total and naïve sample variances. Where homogeneous clustering is assumed, (7.20) reduces to the degrees of freedom given under the random-intercept model by Hedges[416] (p.156), correcting a further typographical error in White and Thomas[42] (p.151).

Table 7.3 Family of Naïve Standard Deviations under Various Model Assumptions

Assumptions			Standardising Metric	
Outcome Variances	**Clustering**	**Sample Sizes**	$s_{den,h}$ **in Sums of Squares**	**Expectation of** $s_{den,h}$
$\sigma^2_{h1} = \sigma^2_{h0} = \sigma^2_{h}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	$\sqrt{\frac{SSE_h}{n_{h1} + n_{h0} - 2}}$	σ_h
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} = n_{h0}$	$\sqrt{\frac{SSE_{h1} + SSE_{h0}}{n_h - 1}}$	$\sqrt{\frac{\sigma^2_{h1} + \sigma^2_{h0}}{2}}$
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	$\sqrt{\frac{SSE_{h1} + SSE_{h1}}{n_{h1} + n_{h0} - 2}}$	$\sqrt{\frac{(n_{h1} - 1)\sigma^2_{h1} + (n_{h0} - 1)\sigma^2_{h0}}{n_{h1} + n_{h0} - 2}}$
$\sigma^2_{bh1} = \sigma^2_{bh0} = \sigma^2_{bh}$ $\sigma^2_{wh1} = \sigma^2_{wh0} = \sigma^2_{wh}$ $\sigma^2_{th1} = \sigma^2_{th0} = \sigma^2_{th}$	$\rho_{h1} = \rho_{h0} = \rho_h$	$k_{h1} = k_{h0} = k_h$ $m_{h1} = m_{h0} = m_h$ So $n_{h1} = n_{h0}$	$\sqrt{\frac{SSW_h + SSB_h}{2(n_h - 1)}}$	$\sqrt{\sigma^2_{th}\left(1 - \frac{(m_h - 1)\rho_h}{n_h - 1}\right)}$
$\sigma^2_{bh0} = 0$ $\sigma^2_{wh1} \neq \sigma^2_{h0}$ $\sigma^2_{th1} \neq \sigma^2_{h0}$	$\rho_{h0} = 0$ $\rho_{h1} \neq 0$	$k_{h0} = n_{h0}$ $m_{h0} = 1$ $n_{h1} \neq n_{h0}$	$\sqrt{\frac{SSW_{h1} + SSB_{h1} + SSE_{h0}}{n_{h1} + n_{h0} - 2}}$	$\sqrt{\sigma^2_{pth\mid\rho_{h0}=0}\left(1 - \frac{(m_{h1} - 1)\rho_{h1}\sigma^2_{th1}}{(n_{h1} - 1)\sigma^2_{th1} + (n_{h0} - 1)\sigma^2_{h0}}\right)}$
$\sigma^2_{bh1} \neq \sigma^2_{bh0}$ $\sigma^2_{wh1} \neq \sigma^2_{wh0}$ $\sigma^2_{th1} \neq \sigma^2_{th0}$	$\rho_{h1} \neq \rho_{h0}$	$k_{h1} \neq k_{h0}$ $m_{h1} \neq m_{h0}$ $n_{h1} \neq n_{h0}$	$\sqrt{\frac{SSW_{h1} + SSB_{h1} + SSW_{h0} + SSB_{h0}}{n_{h1} + n_{h0} - 2}}$	$\sqrt{\sigma^2_{pth}\left(1 - \frac{(m_{h1} - 1)\rho_{h1}\sigma^2_{th1} + (m_{h0} - 1)\rho_{h0}\sigma^2_{th0}}{(n_{h1} - 1)\sigma^2_{th1} + (n_{h0} - 1)\sigma^2_{th0}}\right)}$

Note σ = the population outcome variance, s = its sample estimate, ρ = the population intraclass correlation, k = the number of clusters, m = the cluster size, n = the sample size, h = study, i = arm, b = between, t = total, w = within, e = error, SS = sums of squares

Table 7.4 Sampling Distributions of the Family of Naïve Variances

Assumptions				Degrees of Freedom
Outcome Variances	**Clustering**	**Sample Sizes**	$s^2_{den,h}$	$df_{s^2_{den,h}}$
$\sigma^2_{h1} = \sigma^2_{h0} = \sigma^2_h$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	s^2_h	$n_{h1} + n_{h0} - 2$
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} = n_{h0}$	$s^2_{Huynh,h}$	$\frac{(n_{h1}-1)(n_{h0}-1)(\sigma^2_{h1}+\sigma^2_{h0})^2}{(n_{h0}-1)\sigma^4_{h1}+(n_{h1}-1)\sigma^4_{h0}}$
$\sigma^2_{h1} \neq \sigma^2_{h0}$	$\rho_{h1} = \rho_{h0} = 0$	$n_{h1} \neq n_{h0}$	s^2_{ph}	$\frac{((n_{h1}-1)\sigma^2_{h1}+(n_{h0}-1)\sigma^2_{h0})^2}{(n_{h1}-1)\sigma^4_{h1}+(n_{h0}-1)\sigma^4_{h0}}$
$\sigma^2_{bh1} = \sigma^2_{bh0} = \sigma^2_{bh}$ $\sigma^2_{wh1} = \sigma^2_{wh0} = \sigma^2_{wh}$ $\sigma^2_{th1} = \sigma^2_{th0} = \sigma^2_{th}$	$\rho_{h1} = \rho_{h0} = \rho_h$	$k_{h1} = k_{h0} = k_h$ $m_{h1} = m_{h0} = m_h$ So $n_{h1} = n_{h0}$	s^2_h	$\frac{2((n_h-1)-(m_h-1)\rho_h)^2}{n_h(1+(m_h-1)\rho^2_h)-(1+(m_h-1)\rho_h)^2}$
$\sigma^2_{bh0} = 0$ $\sigma^2_{wh1} \neq \sigma^2_{h0}$ $\sigma^2_{th1} \neq \sigma^2_{h0}$	$\rho_{h0} = 0$ $\rho_{h1} \neq 0$	$k_{h0} = n_{h0}$ $m_{h0} = 1$ $n_{h1} \neq n_{h0}$	s^2_{ph}	$\frac{((1-n_{h1})\sigma^2_{wh1}+(n_{h0}-1)(\rho_{h1}-1)\sigma^2_{h0})^2}{(\rho_{h1}-1)^2\left((n_{h0}-1)^2\sigma^4_{h0}+\left(\frac{(n_{h1}-1)^2}{m^2_{h1}}\left(\frac{m_{h1}-1}{k_{h1}}+\frac{(1+(m_{h1}-1)\rho_{h1})^2}{(k_{h1}-1)(\rho_{h1}-1)^2}\right)\sigma^4_{wh1}\right)\right)}$
$\sigma^2_{bh1} \neq \sigma^2_{bh0}$ $\sigma^2_{wh1} \neq \sigma^2_{wh0}$ $\sigma^2_{th1} \neq \sigma^2_{th0}$	$\rho_{h1} \neq \rho_{h0}$	$k_{h1} \neq k_{h0}$ $m_{h1} \neq m_{h0}$ $n_{h1} \neq n_{h0}$	s^2_{ph}	$\frac{\left(\sum_{i=0}^{1}(n_{hi}-1)\sigma^2_{hi}\right)^2}{\sum_{i=0}^{1}\left(\frac{(n_{hi}-1)^2\sigma^4_{hi}(n_{hi}(1+(m_{hi}-1)\rho^2_{hi})-(1+(m_{hi}-1)\rho_{hi})^2)}{((n_{hi}-1)-(m_{hi}-1)\rho_{hi})^2}\right)}$

Note σ = the population outcome variance, s = its sample estimate, ρ = the population intraclass correlation, k = the number of clusters, m = the cluster size, n = the sample size, h = study, i = arm, b = between, t = total, w = within

7.2.2.3 Sampling Distribution of the Study Estimates

Sampling distributions for SMD estimates all have a similar form. Huynh[41] (pp.4-6) and Hedges[43] (pp.360-362) gave general distributions for biased and unbiased estimators under independence and homoscedasticity, respectively. White and Thomas[42] (p.150) gave a general sampling variance for cluster-randomised trials. These can be extended to give a yet more general sampling distribution, as follows.

Suppose for each of h studies that the mean difference in outcome observed between two randomised groups is

$$\bar{y}_{h1} - \bar{y}_{h0} \sim N\left(\mu_{h1} - \mu_{h0}, \frac{a_{h1}\sigma^2_{h1}}{n_{h1}} + \frac{a_{h0}\sigma^2_{h0}}{n_{h0}}\right) \qquad (7.22)$$

where $\mu = \mu_{h1} - \mu_{h0}$ and $\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})} = \left(\frac{a_{h1}\sigma^2_{h1}}{n_{h1}} + \frac{a_{h0}\sigma^2_{h0}}{n_{h0}}\right)$ are unknown and the a_{hi} denote known constants determined by the choice of metric. Adapting Hedges[43],

$$a_{thi} = 1 + (m_{h1} - 1)\rho_{hi}\text{ , }\ a_{whi} = \frac{1 + (m_{h1} - 1)\rho_{hi}}{1 - \rho_{hi}}\ \text{ or }\ a_{bhi} = \frac{1 + (m_{h1} - 1)\rho_{hi}}{\rho_{hi}} \qquad (7.23)$$

could be used if the SMD denominator is a function of total, within- or between-cluster standard deviations, respectively. Suppose then that $\sigma^2_{den,h} = E\left[s^2_{den,h}\right]/b_h$, where b_h is a known bias in the standardising metric, given by

$$b_{pth} = \left(1 - \frac{(m_{h1} - 1)\rho_{h1}\sigma^2_{th1} + (m_{h0} - 1)\rho_{h0}\sigma^2_{th0}}{(n_{h1} - 1)\sigma^2_{th1} + (n_{h0} - 1)\sigma^2_{th0}}\right) \qquad (7.24)$$

for the pooled total standard deviation. The biased estimator of $\theta_{SMD,h}$ is given by

$$g_{un,h} = \frac{(\bar{y}_{h1} - \bar{y}_{h0})\sqrt{b_h}}{s_{den,h}} \qquad (7.25)$$

where $s^2_{den,h}$ is a quadratic form in normal variates derivable from the study report and the sample means $\bar{y}_{hi}$ and variances s^2_{hi} are mutually independent. Extending Huynh[41] and Hedges[43], $g_{un,h}$ can be re-written as

$$g_{un,h} = \frac{\dfrac{(\bar{y}_{h1}-\bar{y}_{h0})-(\mu_{h1}-\mu_{h0})}{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}} + \dfrac{(\mu_{h1}-\mu_{h0})}{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}}}{\dfrac{s_{den.h}}{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}\sqrt{b_h}}} => \frac{(Z+\Delta_h)\sqrt{b_h}}{\left(\dfrac{s_{den.h}}{\sigma_{den,h}\sqrt{b_h}}\right)\left(\dfrac{\sigma_{den,h}\sqrt{b_h}}{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}}\right)}$$

$$=> \frac{(Z+\Delta_h)}{\left(\dfrac{s_{den.h}}{\sigma_{den,h}\sqrt{b_h}}\right)}\left(\frac{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}}{\sigma_{den,h}}\right) => \frac{(Z+\Delta_h)}{\sqrt{\dfrac{\chi^2_{df_h}}{df_h}}}\sqrt{Var[G_h]} \sim t_{df_h,\varphi_h}\sqrt{Var[G_h]} \qquad (7.26)$$

where t_{df_h,φ_h} is a non-central *t*-distribution with degrees of freedom $df_h = df_{\{s^2_{den,h}\}}$.

The noncentrality parameter is equal to

$$\varphi_h = \Delta_h\sqrt{b_h} = \frac{(\mu_{h1}-\mu_{h0})}{E[s_{den,h}]}\sqrt{\frac{b_h}{Var[G_h]}} = \frac{\dfrac{(\mu_{h1}-\mu_{h0})}{\sigma_{den,h}}}{\sqrt{Var[G_h]}} = \frac{\theta_{SMD,h}}{\sqrt{Var[G_h]}} \qquad (7.27)$$

and

$$Z \sim N(0,1), \Delta_h = \frac{(\mu_{h1}-\mu_{h0})}{\sigma_{(\bar{y}_{h1}-\bar{y}_{h0})}},\ \frac{s_{den.h}}{\sigma_{den,h}\sqrt{b_h}} \sim \sqrt{\frac{\chi^2_{df_h}}{df_h}} \text{ and } Var[G_h] = \frac{\sigma^2_{(\bar{y}_{h1}-\bar{y}_{h0})}}{\sigma^2_{den,h}} \qquad (7.28)$$

This simplifies to Cohen's d where $a_{h1} = a_{h0} = 1$, $b_h = 1$ and $\sigma^2_{h1} = \sigma^2_{h0}$. In which case, $Var[G_h] = 1/n_{h1} + 1/n_{h0}$. It simplifies to Huynh's[41] (p.4) g where $a_{h1} = a_{h0} = 1, b_h = 1$ and $Var[G_h]$ is equal to Huynh's k^2; to White and Thomas'[42] (p.150) g_{un} where $a_{hi} = deff_{hi} = (1+(m_{hi}-1)\rho_h)$, $b_h = 1$ and $\sigma^2_{h1} = \sigma^2_{h0}$ with $Var[G_h] = deff_h(1/n_{h1} + 1/n_{h0})$ or $deff_{h1}/n_{h1} + deff_{h0}/n_{h0}$; and to Hedges'[43] (p.360) D where $a_{h1} = a_{h0} \neq 1, \sigma^2_{h1} = \sigma^2_{h0}$ and $Var[G_h] = a_h/\tilde{N}_h$ where $\tilde{N}_h = (1/n_{h1} + 1/n_{h0})^{-1}$.

It follows from the definition of a non-central *t*-distribution[444] that where $df_h > 2$

$$E[g_{un,h}] = E\left[t_{df_h,\varphi_h}\sqrt{Var[G_h]}\right] = \left(\frac{\sqrt{df_h/2}\,\Gamma[(df_h-1)/2]}{\Gamma[df_h/2]}\right)\varphi_h\sqrt{Var[G_h]} = \frac{\varphi_h\sqrt{Var[G_h]}}{c(df_h)}$$

$$= \frac{\theta_{SMD,h}}{c(df_h)}$$

and that

$$
\begin{aligned}
\sigma^2_{\{g_{un,h}\}} &= \sigma^2_{\{t_{df_h,\varphi_h}\sqrt{Var(G_h)}\}} \\
&= Var[G_h]\left(\left(\frac{df_h}{df_h-2}-\left(\frac{\sqrt{df_h/2}\,\Gamma[(df_h-1)/2]}{\Gamma[df_h/2]}\right)^2\right)\varphi_h^2+\frac{df_h}{df_h-2}\right) \\
&= Var[G_h]\left(\left(\frac{df_h}{df_h-2}-\frac{1}{c(df_h)^2}\right)\left(\frac{\theta_{SMD,h}}{\sqrt{Var[G_h]}}\right)^2+\frac{df_h}{df_h-2}\right) \\
&= Var[G_h]\left(\frac{df_h}{df_h-2}\right)+\theta^2_{SMD,h}\left(\frac{df_h}{df_h-2}-\frac{1}{c(df_h)^2}\right) \\
&= \left(\frac{df_h}{df_h-2}\right)\left(Var[G_h]+\theta^2_{SMD,h}\right)-E[g_{un,h}]^2
\end{aligned}
\qquad (7.29)
$$

as given by Huynh[41] (p.4) and White and Thomas[42] (p.150).

The asymptotic standard error of $g_{un,h}$ is given by

$$
\sigma_{\{g_{un,h}\}}=\sqrt{Var[G_h]+\frac{\theta^2_{SMD,h}}{2df_h}}=\sqrt{Var[G_h]+\frac{c_h\theta^2_{SMD,h}}{2b_h^2}} \qquad (7.30)
$$

This corrects typographical errors in Huynh[41] (p.5) and Hedges[43] (p.361), where $df_h = b_h^2/c_h$ in Hedges[43], due to use of Box’s[445] generalisation of Satterthwaite’s procedure[108] for the degrees of freedom.

The result in (7.29) implies that the unbiased estimator of $\theta_{SMD,h}$ is

$$
g_{adj,h}=c\left(df_{s^2_{den,h}}\right)\left(\frac{(\bar{y}_{h1}-\bar{y}_{h0})\sqrt{b_h}}{s_{den,h}}\right) \qquad (7.31)
$$

This simplifies to Hedges’ g where $a_{h1}=a_{h0}=1$, $b_h=1$ and $\sigma^2_{h1}=\sigma^2_{h0}$. In which case, $Var[G_h]=1/n_{h1}+1/n_{h0}$. It simplifies to Huynh’s[41] (p.6) h where $a_{h1}=a_{h0}=1, b_h=1$ and $Var[G_h]$ is equal to Huynh’s k^2; to White and Thomas’[42] (p.150) g_{adj} where $a_{hi}=deff_{hi}\ =(1+(m_{hi}-1)\rho_h)$, $b_h=1$ and $\sigma^2_{h1}=\sigma^2_{h0}$ with $Var(G_h)=deff_h(1/n_{h1}+1/n_{h0})$ or $deff_{h1}/n_{h1}\ \ +deff_{h0}/n_{h0}$; and to Hedges’[43] (p.362) $DJ(b^2/c)$ where $a_{h1}=a_{h0}\neq 1$, $\sigma^2_{h1}=\sigma^2_{h0}$ and $Var[G_h]\ =a_h/\tilde{N}_h$ where $\tilde{N}_h=(1/n_{h1}+1/n_{h0})^{-1}$.

It follows that

$E\left[g_{adj,h}\right]=c(df_h)E\left[g_{un,h}\right]=\theta_{SMD,h}$ and that

$$\sigma^2_{\{g_{adj,h}\}}=c(df_h)^2\sigma^2_{\{g_{un,h}\}}$$
$$=c(df_h)^2\left(Var[G_h]\left(\frac{df_h}{df_h-2}\right)+\theta^2_{SMD,h}\left(\frac{df_h}{df_h-2}-\frac{1}{c(df_h)^2}\right)\right)$$
$$=\left(c(df_h)^2\left(\frac{df_h}{df_h-2}\right)\left(Var[G_h]+\theta^2_{SMD,h}\right)\right)-\theta^2_{SMD,h} \qquad (7.32)$$

where $df_h>2$, as given by Huynh[41] (p.6) and White and Thomas[42] (p.143).

If $b_h=1$, the variance of $g_{adj,h}$ can be re-written[42] as

$$\sigma^2_{\{g_{adj,h}\}}=\frac{Var[G_h]E\left[U_h^2\right]+\theta^2_{SMD,h}Var[U_h]}{E[U_h]^2} \qquad (7.33)$$

where

$U_h=\frac{\sigma_{den,h}}{s_{den,h}}$, $E[U_h]=\frac{1}{c(df_h)}$, $E\left[U_h^2\right]=\frac{df_h}{df_h-2}$, and $Var[U_h]=\frac{df_h}{df_h-2}-\frac{1}{c(df_h)^2}$

As noted by White and Thomas[42], these properties follow from U_h having a Gamma distribution with shape parameter $df_h/2$ and mean 1. They are equally properties of a noncentral *t*-distribution, in that the raw and central moments of the noncentral *t*-distribution are functions of the noncentrality parameter φ_h whose coefficients are, in turn, functions of the degrees of freedom[446]. Hogben *et al*[446] give coefficients for the first and second central moments and the second raw moment as $E[U_h]$, $Var[U_h]$ and $E\left[U_h^2\right]$, respectively.

White and Thomas[42] (p.150) show that a further adjustment is required to estimate $\sigma^2_{\{g_{adj,h}\}}$, because $\theta^2_{SMD,h}=E\left[g_{adj,h}\right]^2=E\left[g^2_{adj,h}\right]-\sigma^2_{\{g_{adj,h}\}}$. Thus, the standard error of $g_{adj,h}$ is

$$\hat{\sigma}_{\{g_{adj,h}\}}=\sqrt{Var[G]+\frac{g^2_{adj,h}Var[U]}{E\left[U^2\right]}}=\sqrt{Var[G]+g^2_{adj,h}\left(1-\frac{df_h-2}{df_hc(df_h)^2}\right)} \qquad (7.34)$$

As the degrees of freedom are approximate, so is the estimated standard error. The accuracy of the Satterthwaite[108] approximations are dependent on the imprecision with which the component parameters are estimated by their sample counterparts.

Table 7.5 summarises the sampling distributions for the estimators that are considered here for nested therapist designs. It can be seen that clustering and heteroscedasticity affect the study estimate and its standard error via the degrees of freedom, the study estimate via the denominator and its associated bias, where applicable, and finally, the standard error via $Var[G]$.

7.3 One-Step Multilevel Models of the IPD

7.3.1 Data Preparation

When the IPD are available, Goldstein *et al*[44] suggested the following transformation,

$$y'_{hijl} = \frac{y_{hijl} - \bar{y}_{h0}}{s} \qquad (7.35)$$

where y_{hijl} is the outcome for patient l in cluster j of treatment arm i of study h.

They argued that subtracting the control mean $\bar{y}_{h0}$ gives the outcomes within studies a common origin, transforming outcomes in a one-step approach into *differences* from this origin. It was claimed that this is important when the studies use different measurement scales because standardised means, like absolute means, are expected to vary as a function of the study[44] and that it is the differences between standardised means that are assumed to be comparable. In practice, however, subtracting the control mean only affects the estimates of the fixed study effects and is unnecessary[44]. Standardised outcomes could then be used in place of absolute outcomes in the models described in Chapter 6.

As before, the divisor s provides the metric. Goldstein *et al*[44] assumed the population value is known, and equal to the sample estimate, ignoring Hedges'[432, 437] small-sample bias. While this is reasonable for studies that have large effective sample sizes such as those in their example, it will lead to bias otherwise, even if the total sample size is

Table 7.5 Sampling Distributions for Estimators of the Standardised Mean Difference

	Unbiased Estimator $g_{adj,h}$	**Expectation** $\theta_{SMD,h}$	**Estimated Sampling Variance** $\hat{\sigma}^2_{g_{adj,h}}$
Nested Designs (Assuming a Two-level Heteroscedastic Model)			
Pooled Total SD	$c\left(df_{s^2_{pth}}\right)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_{pth}}\right)$	$\frac{\mu_{h1}-\mu_{h0}}{\sigma_{pth}}$	$\frac{\left(\frac{deff_{h1}s^2_{h1}}{n_{h1}}+\frac{deff_{h0}s^2_{h0}}{n_{h0}}\right)}{s^2_{pth}}+g^2_{adj,h}\left(1-\frac{df_{s^2_{pth}}-2}{df_{s^2_{pth}}c\left(df_{s^2_{pth}}\right)^2}\right)$
Pooled Naïve SD	$c\left(df_{s^2_{ph}}\right)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_{ph}}\right)\sqrt{\left(1-\frac{(m_{h1}-1)\sigma^2_{bh1}+(m_{h0}-1)\sigma^2_{bh0}}{(k_{h1}m_{h1}-1)\sigma^2_{th1}+(k_{h0}m_{h0}-1)\sigma^2_{th0}}\right)}$	$\frac{\mu_{h1}-\mu_{h0}}{\sigma_{pth}}$	$\frac{\left(\frac{deff_{h1}s^2_{h1}}{n_{h1}}+\frac{deff_{h0}s^2_{h0}}{n_{h0}}\right)}{s^2_{ph}}+g^2_{adj,h}\left(1-\frac{df_{s^2_{ph}}-2}{df_{s^2_{ph}}c\left(df_{s^2_{ph}}\right)^2}\right)$
Partially Nested Designs (Assuming a Two-level Heteroscedastic Model)			
Pooled Total SD	$c\left(df_{s^2_{pth\mid\rho_{h0}=0}}\right)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_{pth\mid\rho_{h0}=0}}\right)$	$\frac{\mu_{h1}-\mu_{h0}}{\sigma_{pth}}$	$\frac{\left(\frac{deff_{h1}s^2_{h1}}{n_{h1}}+\frac{deff_{h0}s^2_{h0}}{n_{h0}}\right)}{s^2_{pth\mid\rho_{h0}=0}}+g^2_{adj,h}\left(1-\frac{df_{s^2_{pth\mid\rho_{h0}=0}}-2}{df_{s^2_{pth\mid\rho_{h0}=0}}c\left(df_{s^2_{pth\mid\rho_{h0}=0}}\right)^2}\right)$
Pooled Naïve SD	$c\left(df_{s^2_{ph\mid\rho_{h0}=0}}\right)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_{ph\mid\rho_{h0}=0}}\right)\sqrt{\left(1-\frac{(m_{h1}-1)\sigma^2_{bh1}}{(k_{h1}m_{h1}-1)\sigma^2_{th1}+(n_{h0}-1)\sigma^2_{h0}}\right)}$	$\frac{\mu_{h1}-\mu_{h0}}{\sigma_{pth}}$	$\frac{\left(\frac{deff_{h1}s^2_{h1}}{n_{h1}}+\frac{deff_{h0}s^2_{h0}}{n_{h0}}\right)}{s^2_{ph\mid\rho_{h0}=0}}+g^2_{adj,h}\left(1-\frac{df_{s^2_{ph\mid\rho_{h0}=0}}-2}{df_{s^2_{ph\mid\rho_{h0}=0}}c\left(df_{s^2_{ph\mid\rho_{h0}=0}}\right)^2}\right)$
Control SD	$c\left(df_{s^2_{h0}}\right)\left(\frac{\bar{y}_{h1}-\bar{y}_{h0}}{s_{h0}}\right)$	$\frac{\mu_{h1}-\mu_{h0}}{\sigma_{h0}}$	$\frac{deff_{h1}}{n_{h1}}\left(\frac{s^2_{h1}}{s^2_{h0}}\right)+\frac{deff_{h0}}{n_{h0}}+g^2_{adj,h}\left(1-\frac{df_{s^2_{h0}}-2}{df_{s^2_{h0}}c\left(df_{s^2_{h0}}\right)^2}\right)$

large. This can be avoided by first dividing the metric by its correction factor $c(df_h)$, using (7.5) or (7.6) and the degrees of freedom given in Table 7.2 or 7.4. As a result, the one-step models described by Goldstein *et al*[44] give estimates of Cohen's d, but they could be easily adapted to provide estimates of Hedges' g.

In the standard case, s denotes the pooled within-treatment standard deviation s_h. It is either the pooled naïve s_{ph} or control arm standard deviation s_{h0} under the Behrens-Fisher problem, depending on the inference that is of interest to the meta-analyst. If crossed therapist designs are to be combined, the pooled within-treatment standard deviation may be a within, between or total standard deviation. As therapists provide both treatments, the standardisation could operate at either the study or the therapist level. Goldstein *et al*[44] chose to standardise at the cluster-level within studies, adopting a cluster-specific, or conditional, approach. Aggregate meta-analyses implicitly adopt a population-average, or marginal, approach, standardising at the study-level, even in the presence of within-study clustering effects.

Regardless of the availability of IPD, standardising at the cluster-level is not advisable for nested study designs because the clusters relate to only one treatment arm, and a within-cluster metric cannot be pooled across arms within studies. It would be possible to use the within-cluster variance averaged across clusters within studies, but this is a marginal quantity and implies a random-intercept or random-coefficient model for the studies. If the interpretation of an SMD is to be meaningful, its metric should not be confounded with the mean difference within the studies (see Greenland[447] for a similar argument regarding standardised regression coefficients). Suppose, for instance, that the true outcomes were unity in all arms and studies. Rather than equalling zero, the study SMDs would be positive or negative, depending on the respective values of the treatment-specific standard deviations. For this reason, the standardising metric must be common to all arms of a study, especially if there is heteroscedasticity between the arms.

7.3.2 Random-Effects Meta-Analysis Models

Once the data are prepared, the one-step models described in Chapter 6 for absolute mean differences can be applied (see Section 6.3). The choice of model, however, will depend on the characteristics of the studies included and the choice of metric. Starting with the standard case, a random-effects meta-analysis model is appropriate because within-study variance estimates are used to define the study metric. This is given by

$$y'_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(2)} t_l + e_l^{(1)} \qquad (7.36)$$

where α is the mean standardised outcome in the control arm of study 1, x_{hl} and t_l are indicator variables for the other studies and for the treatment arm respectively, β_h and θ are fixed standardised study and treatment effects respectively, and $\tau_{study(l)}^{(2)}$ and $e_l^{(1)}$ are the random treatment effect for study h and the random error for patient l in study h respectively, with $\tau_{study(l)}^{(2)} \sim N\left(0, \tau^2\right)$ and $e_l^{(1)} \sim N(0, 1)$.

Where Glass' SMD is used in the context of the Behrens-Fisher problem, between-arm heteroscedaticity at the patient-level should be taken into account in the meta-analysis as well. The appropriate random-effects model is

$$y'_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(2)} t_l + e_{0l}^{(1)}\left(1 - t_l\right) + e_{1l}^{(1)} t_l \qquad (7.37)$$

where the $e_{0l}^{(1)}$ are now random errors for patient l in the control arm of study h, and the $e_{1l}^{(1)}$ are their treatment arm counterparts. Here, $e_{0l}^{(1)} \sim N(0, 1)$ and $e_{1l}^{(1)} \sim N\left(0, \sigma_{e1}^2\right)$, where $\sigma_{e1}^2 \neq 1$.

If Glass' SMD is instead used for studies with partially nested designs, clustering in the treatment arm should also be taken into account. The meta-analysis model becomes

$$y'_l = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{1therapist(l)}^{(2)} t_l + e_{0l}^{(1)}\left(1 - t_l\right) + e_{1l}^{(1)} t_l \qquad (7.38)$$

where the $u_{1therapist(l)}^{(2)} t_l$ are random effects for therapist j in the treatment arm of study h and $u_{0therapist(l)}^{(2)} \sim N\left(0, \sigma_{u1}^2\right)$. The therapist random effects are assumed equal across studies, resulting in an exchangeable random structure at the study-level within arms.

When the pooled within-treatment standard deviation s_{ph} is used in the context of the Behrens-Fisher problem, Model (7.37) is the appropriate model. Here, $e_{0l}^{(1)} \sim N\left(0, \sigma_{e0}^2\right)$ and $e_{1l}^{(1)} \sim N\left(0, \sigma_{e1}^2\right)$ with $(n_0 - 1)\sigma_{e0}^2 + (n_1 - 1)\sigma_{e1}^2 / n_0 + n_1 - 2 = 1$. If the pooled within-treatment naïve standard deviation s_{ph} is used for nested study designs, Model (7.37) should be extended to allow for within-study clustering, giving

$$y_l' = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{0therapist(l)}^{(2)} \left(1 - t_l\right) + u_{1therapist(l)}^{(2)} t_l + e_{0l}^{(1)} \left(1 - t_l\right) + e_{1l}^{(1)} t_l \quad (7.39)$$

where $u_{1therapist(l)}^{(2)}$ and $u_{0therapist(l)}^{(2)}$ are random effects for therapist j in the treatment and control arms of study h respectively, and $u_{1therapist(l)}^{(2)} \sim N\left(0, \sigma_{u1}^2\right)$, $u_{0therapist(l)}^{(2)} \sim N\left(0, \sigma_{u0}^2\right)$ and $\sigma_{u01} = 0$. The average of the total variances $\left(n_0 - 1\right)\left(\sigma_{u0}^2 + \sigma_{e0}^2\right) + \left(n_1 - 1\right)\left(\sigma_{u1}^2 + \sigma_{e0}^2\right)$ $/\left(n_0 + n_1 - 2\right)$ is equal to one, assuming bias in the naïve standard deviations has been taken into account when preparing the data. Model (7.39) remains appropriate where the pooled total standard deviation s_{pth} is used directly, although this implicitly assumes an unstructured random structure at the study-level. Where the within- and between-cluster variances are homogeneous across arms, Model (7.39) simplifies to

$$y_l' = \alpha + \sum_{h=2}^{H} \beta_h x_{hl} + \theta t_l + \tau_{study(l)}^{(3)} t_l + u_{therapist(l)}^{(2)} + e_l^{(1)} \qquad (7.40)$$

7.4 Application to Counselling in Primary Care

Short-term outcomes relating to the GHQ[314], the Symptom Index[318], the BDI[315, 316, 319, 320] and the HADS depression-subscale[317] were available for 850 patients from seven counselling in primary care trials. Of these, 494 (58%) had been allocated counselling with one of 56 counsellors. Overall, the cluster sizes ranged from 1 to 47, and had a median of 4.5, and an inter-quartile range of 2 to 10.5. Data were available relating to 5 or more patients for 33 of the counsellors. Since all seven of the trials had a partially nested design, some of the complexities were avoided. For simplicity, an exchangeable variance-covariance structure was assumed for the counselling arm in this chapter, consistent with Chapter 5.

7.4.1 Aggregate versus One-Step Meta-Analyses

To reflect common lack of knowledge about cluster size distributions, as in Chapter 6, equal cluster sizes were assumed for all aggregate analyses. The pooled ICC estimate of 0.022 (Chapter 5), based on the non-censored doubly-corrected internal estimates, was used regardless of the model. Second iteration estimates of the standard errors were used throughout for the weights. One-step models were implemented in MLwiN using RIGLS for its flexibility, with the data prepared ignoring small-sample bias from $c(df)$ and bias in the pooled naïve standard deviation pertaining to $\sqrt{b_h}$. Programming details for MLwiN and Stata are given as an appendix (see Section 7.6).

7.4.1.1 Glass' SMD

Table 7.6 summarises the estimates and their standard errors for fixed- and random-effects meta-analyses, progressively relaxing the independence and common variance assumptions within the studies. Glass' pooled SMD and its standard error for the usual aggregate fixed-effects meta-analysis are -0.224 and 0.072. The associated two-sided 95% CI is -0.37 to 0.08 indicating that counselling reduces mental health symptoms in the short-term by an average of about 0.2 standard deviations, although this reduction is not statistically different from zero at the 5% significance level. The equivalent one-step estimate and its standard error are -0.228 and 0.067, with the two-sided 95% CI using the t value, -0.36 to 0.10. As in Chapter 6, the similarity of these results implies that imprecision in aggregate within-study variance estimates is not important in the example under this model. The patient-level variance estimate is not 1.000, however, but 0.870 because this model is inconsistent with the way in which the study metrics were defined. The interpretation of the SMD is therefore problematic under this model.

If within-study clustering is ignored, it can be seen that the one-step random-effects model with heterogeneous patient-level variances is appropriate, as it has an estimate of 1.000 for the patient-level variance in the control arm. Between-study heterogeneity in the SMD estimates has less impact than between-arm heterogeneity in the patient-level variances on the interpretation of the model estimates here. The pooled SMD and its standard error are -0.230 and 0.079. The small observed increase in the standard error is largely due to between-study heterogeneity in the SMD, which is estimated to be 0.011 indicating that about 1% of the total variance is among the studies. The 95%

CI using the t value, given by -0.42 to 0.04, is wider because the degrees of freedom is a function of the number of studies rather than the number of patients in a random-effects meta-analysis. If the aggregate model had been used, the pooled SMD would have been -0.249. Allowing for between-arm heteroscedasticity altered the aggregate estimates but not their one-step counterparts in this example. The explanation for this may be bias in the estimate of τ^2, or it might reflect the need for additional iterations of the weights. In contrast to Chapter 6, the usual and robust standard errors are very similar, but the pattern of results across fixed- and random-effects analyses in Chapter 6 does appear to be replicated.

When within-study clustering is taken into account, the appropriate model remains the random-effects model with heterogeneous patient-level variances. The between-study heterogeneity estimate, $\hat{\tau}^2$, is however zero so its fixed-effects counterpart is equally appropriate in this case. The patient-level variance estimate in the control arm is 0.999 rather than precisely one. The explanation for this is unclear, as the small-sample bias $c(df)$ for Glass' SMD is not affected by clustering in the treatment arm. One probable reason is that the therapist-level variance estimate is biased, since it is 0.066, which is larger than expected. It is given with reference to the variance in the control arm due to the choice of metric, which is inappropriate under a two-level heteroscedastic model at the study-level. Although this needs further investigation, it implies that Glass' SMD is not appropriate for partially nested designs. The impact of model misspecification on the pooled SMD, its standard error or interpretation appears to be unimportant in this example. In fact, Glass' SMD is insensitive to the choice of model, presumably because the between-study heterogeneity is minimal.

7.4.1.2 Pooled Total SMD

Table 7.7 summarises the estimates and their standard errors for the equivalent fixed- and random-effects meta-analyses for the population pooled total SMD. The published meta-analysis (Table 4.2) used a slightly different subset of patients. The equivalent aggregate fixed-effects model with homogeneous patient-level variances in Table 7.7 gives a pooled SMD of -0.259, standard error of 0.072 and 95% CI of -0.40 to -0.12. This is comparable to what was published (SMD=-0.24, 95% CI -0.38 to -0.10), which suggests that excluding 18 patients with missing counsellor identifiers had little impact

Table 7.6 Aggregate versus One-Step Meta-Analyses of Glass' SMD in Outcome between Counselling and Control

AGGREGATE	Ignoring Within-Study Clustering						Allowing for Within-Study Clustering (Internal #1)					
Level 1 Variance	Homogeneous			Heterogeneous			Homogeneous			Heterogeneous		
	Weights		Glass' SMD	Weights		Glass' SMD	Weights		Glass' SMD	Weights		Glass' SMD
Study	% F	% R	(Standard Error)	% F	% R	(Standard Error)	% F	% R	(Standard Error)	% F	% R	(Standard Error)
Boot 1994	13.0	13.2	-0.46 (0.205)	13.8	14.0	-0.46 (0.190)	12.9	13.0	-0.46 (0.214)	13.8	14.0	-0.46 (0.196)
Chilvers 2001	10.6	11.0	0.04 (0.220)	8.1	9.1	0.04 (0.238)	11.3	11.5	0.04 (0.222)	8.5	9.3	0.04 (0.241)
Friedli 1997	14.0	14.1	-0.36 (0.195)	15.9	15.7	-0.36 (0.174)	13.4	13.5	-0.36 (0.207)	15.6	15.5	-0.36 (0.182)
Harvey 1998	13.4	13.6	-0.18 (0.196)	12.9	13.3	-0.18 (0.190)	13.9	14.0	-0.18 (0.201)	13.1	13.4	-0.18 (0.195)
Hemmings 1997	15.0	14.9	-0.07 (0.184)	15.0	14.9	-0.07 (0.175)	13.6	13.6	-0.07 (0.203)	14.1	14.2	-0.07 (0.187)
King 2000	14.9	14.9	-0.48 (0.190)	19.9	18.4	-0.48 (0.158)	15.7	15.6	-0.48 (0.194)	20.9	19.4	-0.48 (0.160)
Simpson 2000	19.1	18.3	-0.06 (0.164)	14.4	14.5	-0.06 (0.179)	19.2	18.7	-0.06 (0.171)	14.0	14.2	-0.06 (0.188)
Fixed: Usual	100.0	100.0	-0.224 (0.072)	100.0	100.0	-0.255 (0.068)	100.0	100.0	-0.225 (0.075)	100.0	100.0	-0.258 (0.071)
Fixed: Robust	100.0	100.0	-0.224 (0.078)	100.0	100.0	-0.255 (0.080)	100.0	100.0	-0.225 (0.079)	100.0	100.0	-0.258 (0.081)
Random: Usual	100.0	100.0	-0.224 (0.078)	100.0	100.0	-0.249 (0.078)	100.0	100.0	-0.225 (0.079)	100.0	100.0	-0.252 (0.079)
Random: Robust	100.0	100.0	-0.224 (0.078)	100.0	100.0	-0.249 (0.079)	100.0	100.0	-0.225 (0.079)	100.0	100.0	-0.252 (0.080)
D-L $\hat{\tau}^2_{\theta_h}$	0.006			0.009			0.003			0.007		

ONE-STEP	Ignoring Within-Study Clustering				Allowing for Within-Study Clustering (Internal #1)			
Level 1 Variance	Homogeneous		Heterogeneous		Homogeneous		Heterogeneous	
Meta-Analysis	Fixed	Random	Fixed	Random	Fixed	Random	Fixed	Random
Intercept	-0.15 (0.099)	-0.11 (0.113)	-0.17 (0.100)	-0.13 (0.116)	-0.10 (0.110)	-0.10 (0.113)	-0.10 (0.118)	-0.10 (0.118)
Chilvers 2001	0.28 (0.136)	0.21 (0.155)	0.31 (0.136)	0.24 (0.158)	0.20 (0.149)	0.19 (0.153)	0.21 (0.156)	0.21 (0.156)
Friedli 1997	0.08 (0.126)	0.06 (0.148)	0.08 (0.125)	0.07 (0.150)	0.05 (0.147)	0.05 (0.150)	0.05 (0.157)	0.05 (0.157)
Harvey 1998	0.18 (0.123)	0.14 (0.150)	0.20 (0.121)	0.15 (0.152)	0.11 (0.143)	0.10 (0.149)	0.10 (0.153)	0.10 (0.153)
Hemmings 1997	0.27 (0.116)	0.20 (0.147)	0.29 (0.114)	0.22 (0.148)	0.18 (0.148)	0.17 (0.152)	0.17 (0.161)	0.17 (0.161)
King 2000	0.03 (0.125)	0.03 (0.144)	0.03 (0.124)	0.03 (0.147)	0.00 (0.140)	0.01 (0.144)	-0.01 (0.149)	-0.01 (0.149)
Simpson 2000	0.24 (0.118)	0.17 (0.138)	0.26 (0.117)	0.19 (0.141)	0.17 (0.133)	0.16 (0.137)	0.18 (0.142)	0.18 (0.142)
Counselling	-0.228 (0.067)	-0.230 (0.078)	-0.229 (0.068)	-0.230 (0.079)	-0.230 (0.075)	-0.230 (0.078)	-0.232 (0.079)	-0.232 (0.079)
$\hat{\tau}^2$		0.012 (0.013)		0.011 (0.012)		0.003 (0.014)		0.000 (0.000)
$\hat{\sigma}^2_v$					0.045 (0.027)	0.043 (0.030)	0.066 (0.031)	0.066 (0.031)
$\hat{\sigma}^2_e$	0.870 (0.042)	0.868 (0.042)	1.003 (0.075)	1.000 (0.075)	0.852 (0.042)	0.852 (0.042)	0.999 (0.075)	0.999 (0.075)
$\hat{\sigma}^2_\xi$			0.775 (0.050)	0.774 (0.050)			0.732 (0.049)	0.732 (0.049)
Counsellor ICC	-	-	-	-	0.050	0.048	0.083	0.083
-2 Log Likelihood	2286.17	2288.48	2279.31	2282.01	2286.30	2286.75	2277.87	2277.87

on these results. At a 5% significance level, the standard SMD is statistically different from zero. The one step counterparts are extremely similar with the pooled SMD being -0.262 and its standard error 0.071. The 95% CI using the t value is also -0.40 to 0.12, so that nothing is gained by using a full-likelihood approach under this model in this example. The patient-level variance estimate is 1.005. As the metric for the one step models was the pooled naïve standard deviation, the appropriate model making standard assumptions is the random-effects model, where the patient-level variance estimate is exactly 1.000.

Under the Behrens-Fisher problem, the appropriate model allows for heteroscedasticity between arms at the patient-level as well. Since its impact on the degrees of freedom and the standard error is a function of the ratio of the sample sizes between arms, the relevant SMD is an extension of Huynh's[41] SMD which allows for a ratio other than one. Unequal patient sample sizes were observed between arms for Boot *et al*[314], Harvey *et al*[317] and Hemmings *et al*[318] all favouring counselling, making this issue pertinent for this example. The pooled SMD and standard error for the aggregate random-effects model with heterogeneous patient-level variances are -0.264 and 0.094, while their one-step counterparts are -0.265 and 0.093. Here it is the average of the patient-level variances i.e. ((494-1)*0.878+(356-1)*1.170)/(850-2) that is equal to 1.000.

The methods described by White and Thomas[42] and Hedges[43] do not apply where the within-cluster variance is not defined for at least one study arm, unequal sample sizes are observed across study arms, or the assumption of between-arm homoscedasticity does not hold for the within- or between-cluster variance components. In the context of a one-step approach, a random-intercept model could be assumed for the studies, but a choice must be made between including patients in the control arms as clusters of size one or as clusters of size n_{h0}. If clusters of size one are used, the within-cluster variance is not defined for the control arm and is estimated solely within the treatment arm. Although the between-cluster variance is available for both arms, it is unlikely to be equal. If clusters of size n_{h0} were used instead, the between-cluster variance is not defined for the control arms. This time, while the within-cluster variance is available in both arms, the number of clusters is unequal, giving disproportionately greater weight to the treatment arm. In neither case is a random-intercept model appropriate.

Table 7.7 Aggregate versus One-Step Meta-Analyses of the Pooled Total SMD in Outcome between Counselling and Control

AGGREGATE	**Ignoring Within-Study Clustering**						**Allowing for Within-Study Clustering (Internal #1)**					
	Homogeneous SD			**Heterogeneous SD**			**Pooled Total SD**			**Pooled Naive SD**		
	Weights		**SMD**	**Weights**		**SMD**	**Weights**		**SMD**	**Weights**		**SMD**
Study	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**	**% F**	**% R**	**(Standard Error)**
Boot 1994	13.0	13.6	-0.54 (0.040)	12.2	13.0	-0.54 (0.046)	12.2	13.0	-0.54 (0.049)	12.2	13.0	-0.54 (0.049)
Chilvers 2001	10.5	11.9	0.03 (0.048)	11.0	12.1	0.03 (0.049)	11.5	12.4	0.03 (0.050)	11.5	12.4	0.03 (0.050)
Friedli 1997	13.9	14.2	-0.42 (0.037)	14.2	14.3	-0.42 (0.039)	13.9	14.1	-0.42 (0.043)	13.9	14.1	-0.42 (0.043)
Harvey 1998	13.5	13.9	-0.20 (0.038)	13.3	13.8	-0.20 (0.041)	13.6	14.0	-0.20 (0.043)	13.6	14.0	-0.20 (0.043)
Hemmings 1997	15.1	14.8	-0.08 (0.034)	12.7	13.4	-0.08 (0.043)	12.0	12.8	-0.08 (0.049)	12.0	12.8	-0.08 (0.049)
King 2000	14.9	14.7	-0.57 (0.036)	16.6	15.8	-0.57 (0.034)	17.4	16.4	-0.57 (0.035)	17.4	16.4	-0.57 (0.035)
Simpson 2000	19.1	16.9	-0.05 (0.027)	19.9	17.5	-0.05 (0.027)	19.3	17.4	-0.05 (0.030)	19.3	17.4	-0.05 (0.030)
Fixed: Usual	100.0	100.0	-0.259 (0.072)	100.0	100.0	-0.264 (0.074)	100.0	100.0	-0.266 (0.076)	100.0	100.0	-0.266 (0.077)
Fixed: Robust	100.0	100.0	-0.259 (0.093)	100.0	100.0	-0.264 (0.096)	100.0	100.0	-0.266 (0.097)	100.0	100.0	-0.266 (0.097)
Random: Usual	100.0	100.0	-0.261 (0.093)	100.0	100.0	-0.264 (0.094)	100.0	100.0	-0.265 (0.094)	100.0	100.0	-0.265 (0.094)
Random: Robust	100.0	100.0	-0.261 (0.093)	100.0	100.0	-0.264 (0.094)	100.0	100.0	-0.265 (0.095)	100.0	100.0	-0.265 (0.095)
D-L $\hat{\tau}^2$	0.057			0.060			0.062			0.062		

ONE-STEP (S$_{ph}$)	**Ignoring Within-Study Clustering**				**Allowing for Within-Study Clustering (Internal #1)**			
Level 1 Variance	**Homogeneous**		**Heterogeneous**		**Homogeneous**		**Heterogeneous**	
Meta-Analysis	**Fixed**	**Random**	**Fixed**	**Random**	**Fixed**	**Random**	**Fixed**	**Random**
Intercept	-0.18 (0.106)	-0.11 (0.127)	-0.20 (0.107)	-0.13 (0.133)	-0.15 (0.113)	-0.11 (0.127)	-0.14 (0.121)	-0.11 (0.133)
Chilvers 2001	0.32 (0.146)	0.20 (0.175)	0.36 (0.146)	0.24 (0.181)	0.27 (0.153)	0.20 (0.174)	0.28 (0.161)	0.22 (0.179)
Friedli 1997	0.09 (0.136)	0.06 (0.168)	0.10 (0.134)	0.07 (0.174)	0.08 (0.148)	0.06 (0.168)	0.07 (0.159)	0.06 (0.176)
Harvey 1998	0.22 (0.132)	0.14 (0.173)	0.24 (0.130)	0.16 (0.178)	0.18 (0.144)	0.13 (0.172)	0.16 (0.153)	0.13 (0.177)
Hemmings 1997	0.31 (0.125)	0.19 (0.170)	0.33 (0.122)	0.21 (0.175)	0.25 (0.144)	0.18 (0.171)	0.24 (0.158)	0.19 (0.178)
King 2000	0.03 (0.134)	0.02 (0.163)	0.03 (0.133)	0.02 (0.169)	0.01 (0.143)	0.02 (0.163)	-0.01 (0.152)	0.00 (0.169)
Simpson 2000	0.28 (0.127)	0.16 (0.157)	0.31 (0.126)	0.19 (0.163)	0.24 (0.135)	0.16 (0.156)	0.24 (0.144)	0.18 (0.162)
Counselling	-0.262 (0.071)	-0.265 (0.092)	-0.263 (0.073)	-0.265 (0.093)	-0.262 (0.076)	-0.265 (0.092)	-0.264 (0.081)	-0.266 (0.092)
$\hat{\tau}^2$		0.024 (0.021)		0.023 (0.019)		0.021 (0.021)		0.015 (0.020)
$\hat{\sigma}_v^2$					0.022 (0.023)	0.012 (0.023)	0.043 (0.027)	0.037 (0.029)
$\hat{\sigma}_e^2$	1.005 (0.049)	1.000 (0.049)	1.177 (0.088)	1.170 (0.088)	0.995 (0.049)	0.995 (0.049)	1.172 (0.088)	1.170 (0.088)
$\hat{\sigma}_\xi^2$			0.881 (0.056)	0.878 (0.056)			0.853 (0.056)	0.854 (0.057)
Counsellor ICC	-	-	-	-	0.022	0.012	0.048	0.042
-2 Log Likelihood	2408.30	2411.02	2399.67	2403.08	2409.02	2410.82	2400.12	2401.69

For illustrative purposes, one-step meta-analysis models were fitted which assumed a random-coefficient model for the studies, including the patients in the control arms as clusters of size one. The pooled SMD and its standard error for the fixed-effects model are extremely similar to the usual estimates. The patient-level variance is 0.995 rather than 1.000 due to bias in the study estimates arising from $c(df)$ and b_h being ignored in the data preparation. Although their combined impact is observable to three decimal places, it is not particularly important in this example. The ICC estimate for this model is, interestingly, 0.022. It may be a coincidence that this is equal to the non-censored doubly-corrected estimate from Chapter 5, but as the variance components included in both models are comparable, it provides support for the methods proposed in Chapter 5. The pooled SMD and its standard error for the random-effects model are identical to the usual estimates. The between-study heterogeneity estimate $\hat{\tau}^2$ is also very similar, which largely accounts for this. The pooled counsellor ICC estimate reduces to 0.012, however, which indicates that some of the between-counsellor variation in outcomes is accounted for by variation between the studies. This is realistic because the treatment protocols varied from study to study.

The recommended model for this metric is the random-effects meta-analysis model for which a two-level heteroscedastic model has been assumed for the studies. Table 7.7 gives the aggregate estimates using the pooled total and naïve standard deviations. It can be seen that the estimates and their standard errors are essentially identical under this model. This implies that using a pooled ICC estimate in the pooled total standard deviation but study-specific ICC estimates in the pooled naïve standard deviation is not important in this example. This is reasonable because no between-study heterogeneity was observed in the ICC estimates in Chapter 5. The pooled SMDs relating to the one-step models are similar for this model, and across all the models using the pooled total and naïve standard deviations. The standard errors relating to the fixed-effects models increase as the complexity of the model increases, while those relating to the random-effects models remain stable regardless of the assumptions made. As in Chapter 6, the D-L estimate of between-study heterogeneity is biased, although the impact of this on the robust standard errors is less pronounced. Under this model, it is the pooled total variance i.e. ((494-1)*(0.854+0.037)+(356-1)*1.170)/(850-2), that should be equal to 1.000. It is equal to 1.008 in this case, due to bias in the study estimates arising from ignoring $c(df)$ and b_h when the data was prepared.

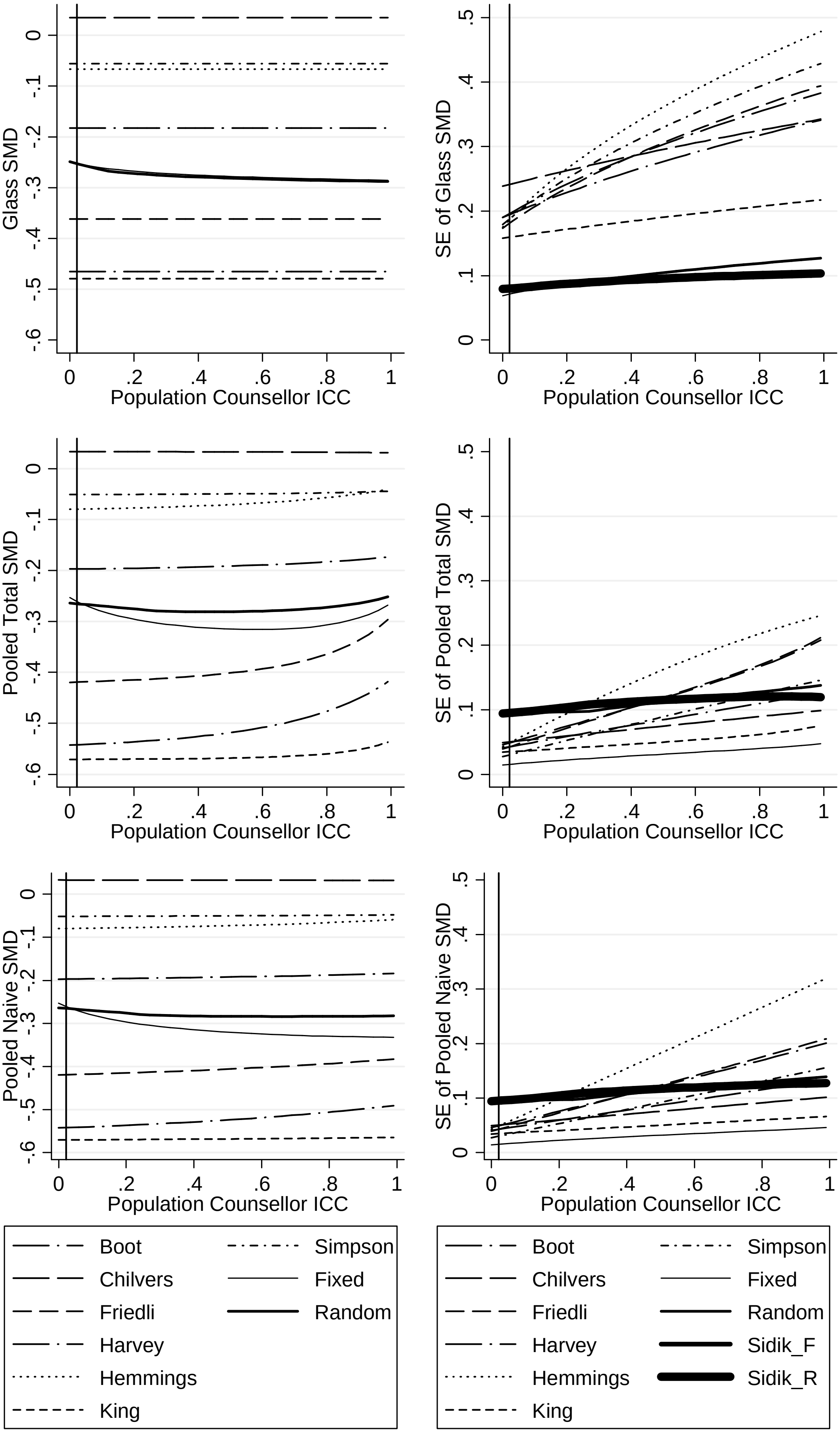
7.4.2 Sensitivity to the Population ICC

The sensitivity of the SMDs using the control, pooled total and pooled naïve standard deviations, and their standard errors, to the population ICC is plotted in Figure 7.2. As in Figure 6.1 for the mean difference (see Section 6.4.2), the dashed lines denote the study-level results, while the solid lines represent the pooled results for the aggregate models allowing for clustering and between-arm heteroscedasticity within the studies. As for mean differences, the study estimates of Glass' SMD are unaffected as the ICC increases. In contrast, the study estimates of the pooled total SMD are pulled towards the pooled estimate, though this is less perceptible in the range of the population ICC expected. This effect is more marked for the pooled total SMD estimate as the degrees of freedom for the pooled total standard deviation reduce at a faster rate than those for the pooled naïve standard deviation, so the impact of the small-sample bias, $c(df)$, is greater in this case. The pooled fixed- and random-effects estimates of Glass' SMD become more extreme at a similar rate as the ICC increases. Pooled random-effects estimates of the pooled total SMD, in contrast, are more stable than their fixed-effects counterparts, providing further support for the conclusion that within-study clustering has less impact on random-effects than on fixed-effects meta-analyses.

The study standard errors are larger for Glass' SMD than for the pooled total or naïve SMDs. This is because the part of the standard error relating to the standardising metric is more important because the effective degrees of freedom are lower here. There is some evidence of between-study heterogeneity in the slope of the standard errors for Glass' SMD but not for the pooled total or naïve SMDs. This arises from use of the control arm standard deviation as the standardising metric. Robust standard errors for the pooled SMDs are fairly stable across the full range of the population ICC. They are also comparable for all three SMD estimates.

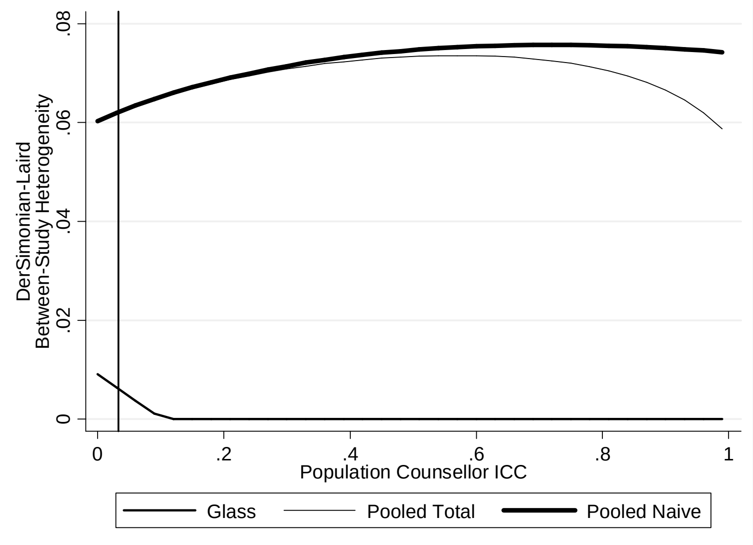
The relationship between the population ICC and the between-study heterogeneity in the SMD is shown below in Figure 7.3. It can be seen that the D-L estimate decreases as the ICC increases for Glass' SMD and is censored at zero when the ICC reaches 0.1. In contrast, the D-L estimate increases for the pooled total and naïve SMDs, until the ICC is in the mid-range, and then decreases again up to its maximum. This difference arises because the total and naïve SMDs are a function of the ICC, while Glass' SMD is not. Even when the ICC is zero the D-L estimate is higher for the total and naïve SMDs

Figure 7.2 Sensitivity Analyses of SMDs and their SEs to the Population ICC



than it is for Glass' SMD. This is because there is more between-study heterogeneity in the outcomes for the counselling arm, when compared to the control arm, as Glass[434] expected.

Figure 7.3 Sensitivity of Between-Study Heterogeneity to the Population ICC



7.5 Discussion

Standardising the mean difference by a standard deviation creates further complexities for the meta-analysis of continuous outcome data from nested therapist designs. The small-sample bias in SMD estimates, and the dependency of the sampling variance on the population parameter, are issues for all meta-analyses of SMDs[432, 437]. Where there is between-arm heteroscedasticity at both the therapist- and patient-levels, the size of the SMD, its small-sample bias, its sampling variance and interpretation depend on the choice of standardising metric. A general approach was described, for an SMD in units of the pooled total standard deviation, which allows the assumptions of independence *and* common variance to be relaxed within studies and the sample size to differ across arms. If the pooled total standard deviation is not available, the pooled naïve standard deviation can be used in its place. Hedges'[437] *g*, Huynh's[41] *h*, White and Thomas'[42] g_{adj}, and Hedges'[43] $DJ\left(b^2/c\right)$ can all be viewed as special cases. This facilitates the pooling of continuous outcomes across studies with diverse designs, because the SMDs have a comparable interpretation. Between-study heterogeneity in the size of the SMD would however be anticipated across study designs, especially if the clustering effect is large.

The example of counselling in primary care was used to illustrate the methods outlined in this chapter. All seven trials had partially nested designs, and there was no evidence of between-study heterogeneity in the counsellor ICC, based on the methods proposed in Chapter 5. The clustering effect was small, as the pooled counsellor ICC was 0.022, the cluster size distribution in the counselling arm was positively skewed, with median 4.5, and a higher number of patients had been allocated counselling compared to no counselling. Consequently, the impact of within-study clustering on the standard error of the pooled SMD estimate was not important. Use of a random-effects meta-analysis model meant the degrees of freedom for testing the hypothesis of no treatment effect were based simply on the number of studies. No allowance was made for imprecision in the counsellor ICC due to the frequentist approach adopted. There was evidence of bias in the D-L estimate of between-study heterogeneity. The usual standard error for the pooled SMD estimate from the random-effects meta-analysis model and the robust standard errors were almost identical however. A similar pattern of results was found if Glass' SMD was used. Further work is needed to evaluate the importance of allowing for treatment-related clustering effects in meta-analyses involving psychotherapy trials more generally.

Glass' SMD, like the absolute mean difference, is not a function of the counsellor ICC under the meta-analysis model assumed for this example. The term within its standard error which relates to the denominator of the SMD is not either, so the dependency of the standard error on the population SMD is unaffected by the size of the within-study clustering effect. The estimate of between-study heterogeneity is lower for Glass' SMD than it is for the pooled total SMD in this example as well. All of these features make it an appealing alternative. It is not clear whether it is appropriate to use it under a two-level heteroscedastic model for the studies, however. Although there is no evidence of between-study heterogeneity in the counsellor ICC for the pooled total SMD consistent with the results in Chapter 5 (see Figure 7.2), there is for Glass' SMD. Misspecification of the variance-covariance structure may, therefore, be responsible for a bias apparent in the counsellor ICC estimate. If this was the explanation, it would simply imply that a more complex model is appropriate when using Glass' SMD for this example.

One advantage of fitting one-step meta-analysis models to the IPD, evident from the results presented here, is that the appropriate model is more obvious. There is a clear relationship between the means by which the data are standardised and the choice of model for the meta-analysis. The impact that model misspecification has on the size or precision of the pooled SMD estimate is also easier to judge on a case-by-case basis. A range of issues were raised by this which deserve further consideration. Firstly, even if no between-study heterogeneity is observed in a SMD, the fixed-effects meta-analysis model is inappropriate because within-study variance estimates are used to define the metric. This is necessary because outcome data is standardised across different scales, standards and study designs, and extends a point Whitehead[102] made about aggregate meta-analyses of absolute mean differences to meta-analyses of SMDs. Secondly, a population-average model is implicitly adopted in aggregate meta-analyses of SMDs. It is arguably necessary for all meta-analyses that involve nested therapist designs too. Further work is needed to evaluate the implications of this in meta-analysis models including within-study clustering effects, extending the work of Bohning *et al*[431] and Viechtbauer[422] to a more general setting. Thirdly, the assumption of a common origin across studies is less justified when data are obtained from different scales, standards or study designs. If a random study intercept were included in a one-step model of the IPD, correlation between heterogeneity in the SMD and its origin can be estimated. While this correlation is a nuisance for aggregate meta-analyses, it might be of interest in one-step meta-analyses.

7.6 Appendix: Programming Code for One-Step Models

STATA VERSION 11

The data were prepared for a dataset IPD_wide.dta with variables study, treat, outcome, n1, n0, m1, s1_sq, s0_sq and icc using the Stata code

```
use IPD_wide.dta, clear
gen pv=((n1-1)*s1_sq+(n0-1)*s0_sq)/(n1+n0-2)          ** POOLED NAÏVE VARIANCE
gen st1_sq=s1_sq/(1-((m1-1)*icc/(n1-1)))             ** COUNSELLING ARM TOTAL VARIANCE
gen ptv=((n1-1)*st1_sq+(n0-1)*s0_sq)/(n1+n0-2)       ** POOLED TOTAL VARIANCE
gen ss_s0_sq=outcome/sqrt(s0_sq)                     ** STANDARDISED SCORES
gen ss_pv=outcome/sqrt(pv)
gen ss_ptv=outcome/sqrt(ptv)
gen o_s0_sq=.                                        ** STUDY ORIGINS
gen o_pv=.
gen o_ptv=.
```

```
levelsof study, local(slist)
foreach study in `slist' {
        qui sum ss_s0_sq if treat==0 & study==`study'
        replace o_s0_sq=r(mean) if study==`study'
        qui sum ss_pv if treat==0 & study==`study'
        replace o_pv=r(mean) if study==`study'
        qui sum ss_ptv if treat==0 & study==`study'
        replace o_ptv=r(mean) if study==`study'
}
gen outcome_glass=ss_s0_sq-o_s0_sq                    ** TRANSFORMED OUTCOMES
gen outcome_pnaive=ss_pv-o_pv
gen outcome_ptotal=ss_ptv-o_ptv
```

A random-effects meta-analysis (model 7.37) for Glass' SMD under the Behrens-Fisher problem can be fitted using

```
xi: xtmixed outcome_glass i.study i.treat || study: treat, nocons resid(ind, by(treat))
```

If clustering is taken into account in the treatment arm (model 7.38), this becomes

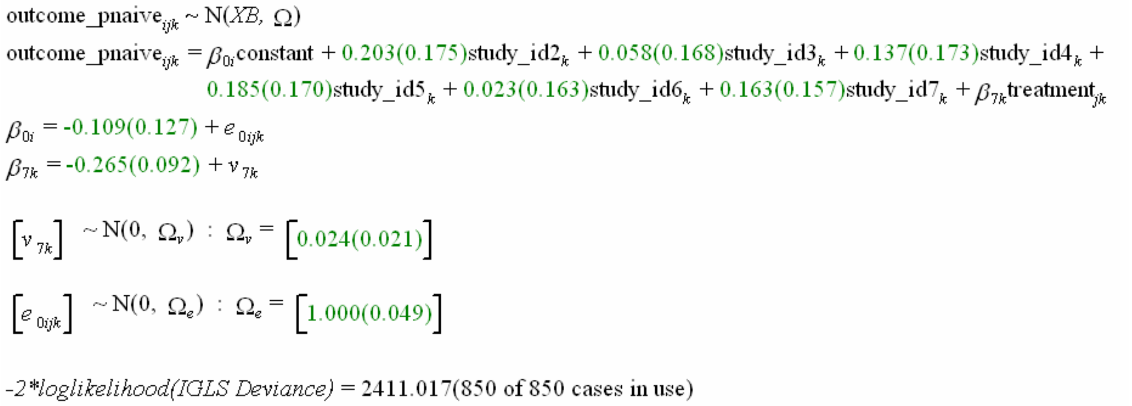
```
xi: xtmixed outcome_glass i.study i.treat || study: treat, nocons || t_id: treat, nocons
resid(ind, by(treat))
```

where t_id is the therapist identifier. The outcome outcome_glass can be replaced by outcome_pnaive or outcome_ptotal for the pooled naïve or pooled total SMDs. Model 7.40 is given by

```
xi: xtmixed outcome_pnaive i.study i.treat || study: treat, nocons || t_id:
```

MLwiN VERSION 2.02

A dataset was imported starting with variables study_id, t_id, p_id identifying the study, cluster (i.e. counsellor or control patient), and patient, followed by indicator variables study_id2, study_id3, study_id4, study_id5, study_id6, study_id7 for Chilvers 2001, Friedli 1997, Harvey 1998, Hemmings 1997, King 2000 and then Simpson 2000, treatment for counselling, control for no counselling, outcome_glass, outcome_pnaive, outcome_ptotal for the outcomes, and constant for a column of ones. The data were already sorted on study_id, t_id, and p_id and had been reduced to complete cases. Once in MLwiN, the RIGLS option under *Equations* was used, and the *Equations* under *Model* was used to open an interactive window. The outcome was specified and the levels. The standard random-effects meta-analysis model (7.36) was fitted for outcome_pnaive as follows:



$$\text{outcome\_pnaive}_{ijk} \sim \mathrm{N}(XB,\ \Omega)$$

$$\text{outcome\_pnaive}_{ijk} = \beta_{0i}\text{constant} + 0.203(0.175)\text{study\_id2}_k + 0.058(0.168)\text{study\_id3}_k + 0.137(0.173)\text{study\_id4}_k + 0.185(0.170)\text{study\_id5}_k + 0.023(0.163)\text{study\_id6}_k + 0.163(0.157)\text{study\_id7}_k + \beta_{7k}\text{treatment}_{jk}$$

$$\beta_{0i} = -0.109(0.127) + e_{0ijk}$$

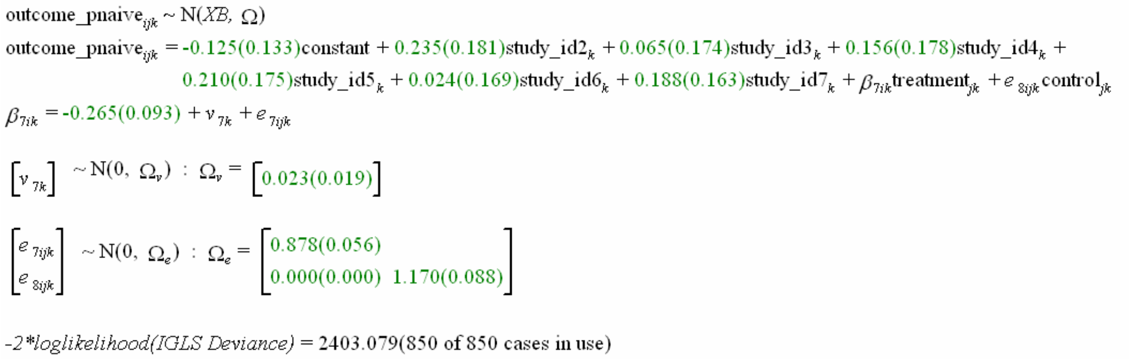
$$\beta_{7k} = -0.265(0.092) + v_{7k}$$

$$\begin{bmatrix} v_{7k} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_v)\ :\ \Omega_v = \begin{bmatrix} 0.024(0.021) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 1.000(0.049) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2411.017(850 of 850 cases in use)

Under the Behrens-Fisher problem (see model 7.37) this becomes



$$\text{outcome\_pnaive}_{ijk} \sim \mathrm{N}(XB,\ \Omega)$$

$$\text{outcome\_pnaive}_{ijk} = -0.125(0.133)\text{constant} + 0.235(0.181)\text{study\_id2}_k + 0.065(0.174)\text{study\_id3}_k + 0.156(0.178)\text{study\_id4}_k + 0.210(0.175)\text{study\_id5}_k + 0.024(0.169)\text{study\_id6}_k + 0.188(0.163)\text{study\_id7}_k + \beta_{7ik}\text{treatment}_{jk} + e_{8ijk}\text{control}_{jk}$$

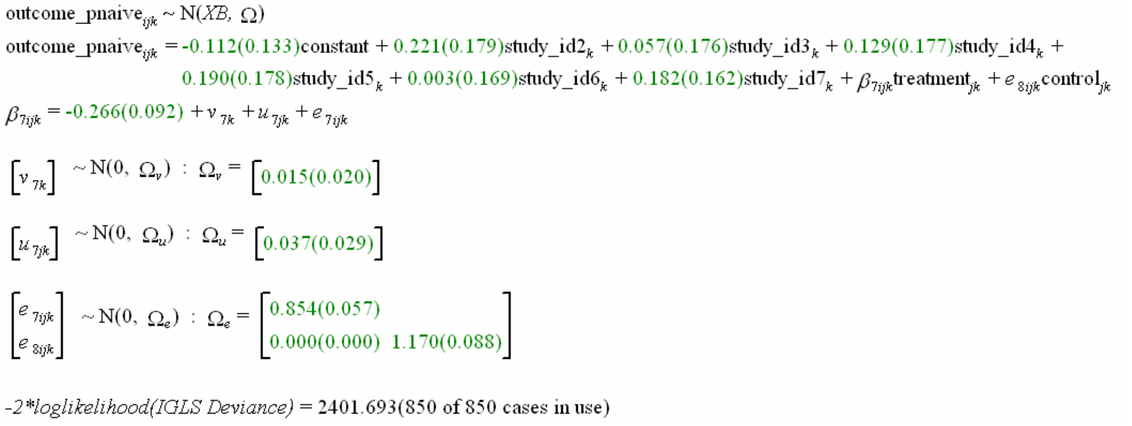
$$\beta_{7ik} = -0.265(0.093) + v_{7k} + e_{7ijk}$$

$$\begin{bmatrix} v_{7k} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_v)\ :\ \Omega_v = \begin{bmatrix} 0.023(0.019) \end{bmatrix}$$

$$\begin{bmatrix} e_{7ijk} \\ e_{8ijk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.878(0.056) & \\ 0.000(0.000) & 1.170(0.088) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2403.079(850 of 850 cases in use)

Allowing for clustering in the treatment arm (see model 7.38) it is



$$\text{outcome\_pnaive}_{ijk} \sim \mathrm{N}(XB,\ \Omega)$$

$$\text{outcome\_pnaive}_{ijk} = -0.112(0.133)\text{constant} + 0.221(0.179)\text{study\_id2}_k + 0.057(0.176)\text{study\_id3}_k + 0.129(0.177)\text{study\_id4}_k + 0.190(0.178)\text{study\_id5}_k + 0.003(0.169)\text{study\_id6}_k + 0.182(0.162)\text{study\_id7}_k + \beta_{7ijk}\text{treatment}_{jk} + e_{8ijk}\text{control}_{jk}$$

$$\beta_{7ijk} = -0.266(0.092) + v_{7k} + u_{7jk} + e_{7ijk}$$

$$\begin{bmatrix} v_{7k} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_v)\ :\ \Omega_v = \begin{bmatrix} 0.015(0.020) \end{bmatrix}$$

$$\begin{bmatrix} u_{7jk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.037(0.029) \end{bmatrix}$$

$$\begin{bmatrix} e_{7ijk} \\ e_{8ijk} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.854(0.057) & \\ 0.000(0.000) & 1.170(0.088) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 2401.693(850 of 850 cases in use)

8 DISCUSSION

The objective of this thesis was to develop a conceptual framework for understanding the role of therapists in psychotherapy trial designs and to review, adapt, illustrate and compare methods for meta-analysing trials involving psychotherapy. Chapter 1 set out the rationale and scope of the thesis. Chapter 2 described a framework for considering therapist variation, using the broad concepts of precision, internal and external validity to outline the implications of nesting of patients within therapists for randomised trials of psychotherapy. Chapter 3 systematically reviewed Cochrane reviews of comparative studies involving psychotherapy, exploring the range, complexity and recognition of issues arising from the multilevel aspects of their designs. It was clear that reviewers were aware that therapist variation had implications but were unfamiliar with those for precision due to the clustering effects. Chapter 4 introduced the counselling in primary care meta-analysis that was used to illustrate the methods described in Chapters 5, 6 and 7. Chapter 5 adapted methods described by Blitstein *et al*[38] for the meta-analysis of ICC estimates, comparing several methods for reducing bias. Chapter 6 extended the methods outlined by Kwong and Higgins[39] for meta-analysing absolute mean differences. And finally Chapter 7 integrated the methods described by Huynh[41], White and Thomas[42] and Hedges[43] for aggregate data, and extended those described by Goldstein *et al*[44], for the meta-analysis of standardised mean differences, using a one-step multilevel model for nested therapist designs. The novel aspects of this work are outlined in this chapter.

8.1 A Conceptual Framework

Psychotherapy research has developed its own methodological literature, published in books, book chapters, and subject-specific journals, separated from the statistical and trial methodology literature. This dates back at least as far as the 1952 special issue of *Journal of Clinical Psychology*[448-451]. It is clear from this that the therapist is central to psychotherapy and to psychotherapy research designs. While the complexities inherent in conducting research in this area are widely recognised and debated, a methodology that adequately addresses them has been slower to develop. One reason for this is the lack of a clear, and sufficiently broad, conceptual framework. In 1966, Kiesler[45] argued that

> "One of the unfortunate effects of the prolific and disorganized psychotherapy research

> literature is that a clear-cut, methodologically sophisticated, and sufficiently general paradigm which could guide investigations in the area has not emerged. Perhaps this is an unavoidable state of affairs in a new area of research. Yet a perusal of this literature indicates that most of the basic considerations necessary for a general paradigm have appeared, albeit in many cases parenthetically, at some place or another. But to date no one has attempted to integrate empirical findings and methodological concerns in a way that might lead to a useful research paradigm. This lack of integration of the paradigm ingredients has minimized their impact on investigators in the area." (p. 110)

Regrettably, this remains largely the case 40 years later. One explanation is that, over the intervening period, generic methodologists, among them statisticians, have tended to be more familiar with the statistical and mainstream literature than with the subject specific one. It is only recently that issues concerning therapist variation were raised in the statistical literature[81] with no reference made here to the psychotherapy literature. Since 1999, attention has focused on the implications for precision of treatment effect estimates. Reading this one might be forgiven for linking the role of the therapist in randomised trials, from a statistical perspective, simply to the size and precision of the associated clustering effect. The therapist, however, is more fundamentally part of a complex or multi-component intervention. They are on the causal pathway from the psychotherapy to the patient. It is their joint status as a potential treatment factor and experimental unit which makes them central to the design and analysis of randomised trials of psychotherapy. Characteristics of the therapist delivering the intervention are important regardless of the size of the clustering effect because of the implications for internal and external validity. They are of specific interest as the predictors of therapist variation could be used to inform the selection and training of therapists.

The conceptual framework developed provides one basis for understanding the nature and impact of performance bias, and by extension, selection, detection and attrition biases in all clinical trials. By tackling the complexities around the role of therapists in psychotherapy trials, a general research paradigm has begun to emerge, with the potential to guide the design and analysis of complex intervention trials in the future, when more fully developed. This should contribute to a clearer understanding of the methodology used in drug trials and epidemiological studies. The broad concept of multiple levels of experimental units joins longitudinal, multicentre, cluster-randomised and crossover trials together, since each has a *multilevel trial design*. The inability to randomise centres, time or clusters to patients leads to cluster sampling of outcomes, necessitating an observational component, which perhaps deserves greater attention. The combination of experimental and observational design aspects in randomised trials has implications for their analysis and interpretation. Clarity is needed to ensure they follow directly from, and are thus appropriate to, the trial design.

The concept of multiple treatment variables characterising complex interventions, with some fixed and others random, some categorical, others continuous, has the potential for facilitating greater understanding of the components of complex interventions, how they interact, which are important, to what extent, and for whom. This brings what is currently referred to as *process research* within the remit of randomised trials enabling a more complete evaluation of the causal effects of multi-component interventions. In this thesis, two components have been considered – the therapeutic approach and the therapist. Both are categorical, the former is fixed and the latter random. This provides a relatively simple illustration of a more general paradigm. Going back to the example comparisons given in Figure 2.1, it is clear that what is being proposed is an extension to factorial trial designs. Currently, the majority of psychotherapy trials are incomplete factorial trials of packages of therapeutic approaches and therapist characteristics. The PACE trial[3] discussed in Chapter 1 is typical of this. It is consistent with the evaluation of the effectiveness of complex interventions, implying a pragmatic research question. This could be viewed as a little premature if a detailed understanding does not exist of the causal effects of the components of the therapeutic approach, and their interaction with therapist characteristics. At present, earlier-phase psychotherapy trials tend to be smaller versions and do not provide a sufficient basis for determining the optimal form of a complex intervention for taking forward to a large-scale definitive trial. This is why psychotherapy researchers have found the drug metaphor and associated trial designs uncomfortable for complex interventions.

Integrating the relevant psychotherapy and statistical literatures on therapist variation has proved fruitful. Initial discussions with psychiatrists, psychologists and statisticians working in the field helped to locate different sections of this literature. The references and more focused searches identified the remainder. Although extensive, the approach taken was not fully systematic, based on the standards of Cochrane reviews. It served to provide the ingredients necessary to develop a conceptual framework however. The systematic methodological review of Cochrane reviews described in Chapter 3 provided a means of assessing how adequate and complete this framework was. This generated in its turn an overview of further areas of methodological research currently needed. It is clear in hindsight that this component of the thesis is a form of qualitative research.

Its potential as a means for improving the speed with which important methodological issues are addressed is also clear. Over time, as the relevant methodological literature becomes increasingly vast and disparate, its implementation by researchers in the field requires them to be aware of more, to synthesise more, and to have a greater level of expertise. For the same reasons as it became necessary to summarise clinical research in Cochrane reviews, now is perhaps time to do so on a larger scale for methodological research. Systematic methodological reviews could then feed into the guidance given to researchers on reporting of primary research studies, and more generally in courses and textbooks. Regular updates of this guidance might then help them keep abreast of developments, and ultimately improve the standard of research and patient care. This has the potential to avoid unnecessary duplication of effort and to help to set priorities for future methodological research.

8.2 Therapist Variation in Meta-Analyses of Psychotherapy

Traditionally, meta-analyses of treatment effects provided by randomised trials involve two levels: one represents the trials and the other the patients in the trials. As in other multilevel situations, implications of both levels should be considered for the precision, internal and external validity of treatment effect estimates. Precision is affected by the number of trials and patients-per-trial, and the relationships among treatments, trials, and patients. Internal validity is affected by the nature of the allocations of treatments to trials, trials to patients, and treatments to patients. External validity is then affected by the selection of trials, and of patients within trials. As in many psychotherapy trials, only the allocation of treatments to patients is random in meta-analyses, and then only within the trials. The importance of investigating treatment-by-covariate interactions is hence raised above that of an exploratory analysis. Heterogeneity in the patient-level variance may also be expected between trials, unless the patient eligibility criteria and sample characteristics are identical across the trials. Accordingly, greater attention is needed to the observational aspects of meta-analyses as well.

Additional levels in meta-analyses of psychotherapy trials have further implications for the precision, internal and external validity of treatment effect estimates. Nevertheless it was clear from the systematic methodological review described in Chapter 3 that the precision implications of therapist variation had not been considered within any of the relevant reviews published in Issue 1, 2007, of the *Cochrane Database of Systematic*

Reviews. There is also no published methodological guidance, although a paper was in preparation and a draft of this was kindly shared[39]. The methods proposed in Chapters 5, 6 and 7 can be viewed as extensions of this. It has been argued here that between-arm heteroscedasticity at the therapist- and patient-levels in trials affects the choice of an appropriate model for meta-analyses. The size of the therapist and patient samples and a current lack of published therapist ICC estimates add other complexities. Use of individual-patient-data to conduct meta-analyses is also rare, necessitating adoption of an aggregate approach, based on the summary statistics and standard errors available in published reports. In part in response, Kwong and Higgins[39] restricted their work to aggregate methods. They derived the sampling distribution of a mean difference under a two-level heteroscedastic model, and suggested a method for obtaining internal ICC estimates from outcome variances often reported. They proposed a general approach for meta-analyses of odds ratios, absolute and standardised mean differences. Finally, they assessed the sensitivity of conclusions to assumptions about the underlying ICC.

Meta-analyses of the individual-patient-data tend to have been justified on the basis of problems with meta-analyses of aggregate data obtained from published reports[152, 428, 452-456]. Yet, the distinction between IPD and aggregate approaches confounds whether data for the meta-analysis is obtained from the published reports or original datasets with whether an aggregate or one-step approach is taken to the analysis of this data. The limitations of using published reports are predominantly practical, while those that relate to use of an aggregate approach are predominantly statistical. In this context, it is expected that therapist ICC estimates will not be available in published reports, and that the analyses presented will ignore treatment-related clustering effects associated with therapists. So, summary statistics and appropriate standard errors taking account of clustering, which are necessary for an aggregate meta-analysis, are expected to be missing from published reports. This data could be imputed from other sources, but if this is done the same principles apply here as to all forms of missing data. Uncertainty in the value of these estimates should be formally taken into account. Little is currently known about the size or predictors of therapist ICC estimates. Thus, one advantage of collecting the IPD for meta-analyses of psychotherapy trials is that doing so minimises the impact of uncertainty on the precision of the treatment effect estimate. Given the imprecision of ICC estimates, the additional resource required might be an acceptable price to pay.

The distinction between aggregate and one-step meta-analyses mirrors that of cluster-level and individual-level analyses of cluster-randomised trials. However, in the context of an aggregate meta-analysis, it is common practice to assume the estimated weights are known and the number of trials and patients sampled is infinite[102, 457]. It is not entirely clear why this is the case. It does reduce the apparent complexity of the analyses, but at a cost. It makes it more difficult for researchers to generalise their understanding of multilevel analyses gained from other areas, and less easy for them to appropriately assess the validity of model assumptions in their circumstances. As it is the principal approach in use, textbooks, such as Whitehead[102], include equivalent one-step models rather than aggregate counterparts to them. Comparison of the two approaches was helpful here for clarifying their strengths and limitations.

One potential limitation of the usual aggregate approach, discussed in Chapter 6, is its failure to fully allow for the uncertainty present when testing hypotheses regarding the treatment effect. The impact is more apparent if the number of studies, therapists or patients is small and the corresponding design effects are large. Cornfield[337] described clustering penalties for the standard error and effective degrees of freedom. Kwong and Higgins[39] considered the implications for the standard error but not the degrees of freedom. These were considered here in the context of meta-analyses of standardised mean differences because they are important in determining the extent of Hedges'[432, 437] small-sample bias and the standard error of a standardised mean difference. They are also important for the meta-analysis of absolute mean differences where a fixed-effects meta-analysis model is fitted in the presence of within-study clustering. If a random-effects meta-analysis model is adopted, the number of studies determines the degrees of freedom, because the studies represent the highest level in the model. The number of therapists is thus less important for meta-analyses than it is for randomised trials. Its importance instead lies in the precision of the estimated weight given to each study. This presumably depends on the sampling distribution of the standard error of the treatment effect. Since this is unknown and likely to be a function of the therapist ICC,use of a robust sandwich estimator offered a compromise. The methods described by Sidik and Jonkman[40] were therefore adapted.

An issue that became apparent, when comparing the estimates and standard errors of one-step models with those based on the robust sandwich estimator, was bias in a D-L estimate of between-study heterogeneity of the treatment effect (see Chapter 6). This appeared to bias the robust sandwich estimator, when it was used in conjunction with a random-effects meta-analysis. Component-wise estimation of the variance terms and failure to account for their imprecision provides a likely reason[422]. However the impact of bias on a robust sandwich estimator deserves further consideration, since it implies model misspecification is less important than the use of unbiased estimates. A second issue, highlighted by one-step meta-analysis models of standardised mean differences, is a relationship between the data, or the method of data handling, and the choice of model for the meta-analysis. It was apparent that fixed-effects meta-analysis models were not generally appropriate when obtaining data from published study reports. This is because they ignore the cluster sampling of patients within studies. Similarly, bias is known to result from between-study heterogeneity in the patient-level variance[422]. As this is expected, due to non-random allocation of studies to patients, it can be argued that a random-effects meta-analysis model is appropriate even if there is no between-study heterogeneity in the treatment effect.

Kwong and Higgins[39] generalised an approach proposed for absolute mean differences to odds ratios and to standardised mean differences. It has been shown that methods for ICCs, absolute and standardised mean differences share common features, but are also quite specific. Use of a population-average or cluster-specific meta-analysis model becomes an issue deserving of greater attention where the summary measure is not a mean difference. This was discussed in Chapter 7 in relation to the use of therapist- or study-level standardising metrics and the need for a common metric for all arms in the studies. It is not clear at this stage what should be done for odds ratios, relative risks, risk differences, or indeed hazard ratios.

The most influential papers on therapist variation in the psychotherapy field are those of Crits-Christoph and colleagues[80, 363] but the statistical methods used in these papers are unsophisticated. In contrast, the idea of pooling therapist ICC estimates to explore predictors and to inform methods for minimising the clustering penalties in early-phase trials is reasonably advanced, even by today’s standards. Baldwin *et al* were preparing an extension to these papers and kindly shared a draft[364]. The statistical methods used reflect recent developments. These therefore served as the starting point in Chapter 5. Psychotherapy trials, unlike cluster-randomised trials, often constitute the majority, if not all, the trials in relevant meta-analyses, as was seen in Chapter 3. As such, pooling of therapist ICC estimates and exploration of their predictors is an additional analysis of interest in this setting. As with any meta-regression, a large number of studies are required for adequate precision. The number of predictors which can be considered is further limited by the precision of the ICC estimates within the trials. As the chance of obtaining negative ICC estimates is higher in smaller trials, this has implications for the complexity of the assumed variance-covariance matrix in meta-analyses of early-phase psychotherapy trials, as observed in Chapter 6. An unstructured variance-covariance matrix led to computational problems, while an exchangeable one might be unrealistic. Nevertheless the presence of between-study heterogeneity in the ICCs is an important consideration. It was evident from the illustrative example that, where ICCs are closely matched, much of the heterogeneity in estimates can be attributed simply to sampling variation. This reflects the level of imprecision of these estimates in the counselling in primary care trials seen in Chapter 5. Since the sample size is limited in meta-analyses of randomised trials, large naturalistic databases may initially provide a better setting for exploring predictors. The relative ease of fitting complex models highlights one of the main practical reasons for using the IPD.

One dilemma faced by the meta-analyst in this setting is the interpretation and choice of metric for standardised mean differences in the presence of clustering. The options, being numerous, fall into two broad categories, relating to specific or pooled standard deviations. It is arguable which is easier to interpret and it will probably depend on the circumstances. For the example of counselling in primary care, the small size of the clustering effect makes the pooled total standard deviation a preferred metric over the control arm standard deviation, as its interpretation is similar to Hedges' *g*. The need for a common metric across treatment arms defined on a study-by-study basis implies a treatment-specific unstructured variance-covariance matrix, if the metric is based on a specific standard deviation. In the case of pooled metrics, the averaging of two or more population standard deviations makes the random structure somewhat contrived. In a one step meta-analysis, the numerator and denominator of an ICC are separately estimated. Since the denominator depends on the choice of metric, the ICC varies as a function of the metric even when the model is correctly specified.

8.3 Limitations and Future Work

It is clear that this is a fairly untouched area in need of further exploration. Limitations of the thesis are linked to the possibilities for future work. For example, the conceptual framework could be elaborated to include situations in which there are more than two levels, such as repeated measurements in therapist designs and trials of group-based interventions. Further consideration could be given to use of multi-tiered experimental designs[138] in early-phase psychotherapy trials, reflecting possible interest in therapeutic approaches *per se* rather than their combination with specific therapist characteristics. Similarly, the advantages of crossed designs for early-phase trials could be considered further in relation to learning curves, as might the possibility of evaluating the optimal levels of multiple interacting treatment components using response surface designs[458]. The aim of this work would be respond more fully to Kazdin's[26] challenge of using experimental manipulations to explore both common and specific factors contributing to psychotherapy, providing a greater understanding of the individual causal effects of a complex intervention. A broader investigation of issues arising in the context of other complex interventions, such as surgery and physiotherapy, might also be carried out.

Further work is needed to evaluate the meta-analysis methods proposed here. As was discussed in Chapter 5, simulation work is needed to assess the extent of residual bias in the study and pooled ICC estimates from approximations for ANOVA estimates and average cluster sizes. This work could also include an assessment of the adequacy of Fisher's[377] approximate standard error and the D-L estimate of between-study heterogeneity in this context. As was discussed in Chapter 6, simulation work is also needed to evaluate the use of Sidik and Jonkman's[40] robust 'sandwich' estimator in an aggregate meta-analysis of clustered data, with the aim of elucidating the impact of bias in the between-study variance estimate. This might also include an assessment of the scenarios with which allowance should be made for imprecision of the estimated weights. The role of imprecision in the clustering effect might also be explored within a Bayesian framework, adapting existing methods[459]. Further work is also needed in relation to Chapter 7 to assess the impact of unequal cluster sizes on the proposed methods. It might be possible to do this within a sensitivity analysis, where relevant information is unavailable. Finally extensions are needed for meta-analyses involving crossed designs and for binary and survival outcomes. Methods proposed for repeated measures[149, 151, 152, 460] could be extended to allow for other sources of clustering, such as therapist variation.

References

1. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: The new Medical Research Council guidance. *British Medical Journal* 2008; **337**:979-983.
2. Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and Petticrew, M. Developing and Evaluating Complex Interventions: New Guidance. 2008.
Ref Type: Report
3. White PD, Sharpe MC, Chalder T, DeCesare JC, Walwyn R. Protocol for the PACE trial: A randomised controlled trial of adaptive pacing, cognitive behaviour therapy, and graded exercise as supplements to standardised specialist medical care versus standardised specialist medical care alone for patients with the chronic fatigue syndrome/ myalgic encephalomyelitis or encephalopathy. *BMC Neurology 7, Article Number: 6* 2007.
4. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**(1):103-115.
5. Richardson P. ABC of mental health. Psychological treatments. *British Medical Journal* 1997; **315**(7110):733-735.
6. Hajek P. Current issues in behavioral and pharmacological approaches to smoking cessation. *Addictive Behaviors* 1996; **21**:699-707.
7. Wampold BE. *The Great Psychotherapy Debate: Models, methods, and findings*. Lawrence Erlbaum Associates Publishers: Mahwah, NJ, US, 2001.
8. Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* 2000; **19**(8):1115-1139.
9. Andersen PK, Klein JP, Zhang M-J. Testing for centre effects in multi-centre survival studies: A Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* 1999; **18**(12):1489-1500.
10. Anello C, O'Neill RT, Dubey S. Multicentre trials: A US regulatory perspective. *Statistical Methods in Medical Research* 2005; **14**(3):303-318.
11. Dragalin V, Fedorov V. Design of multi-centre trials with binary response. *Statistics in Medicine* 2006; **25**(16):2701-2719.
12. Fedorov V, Jones B. The design of multicentre trials. *Statistical Methods in Medical Research* 2005; **14**(3):205-248.
13. Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine* 2004; **23**(3):369-388.
14. Gould AL, Freeman PR. Multi-centre trial analysis revisited. *Statistics in Medicine* 1998; **17**(15-16).
15. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine* 1998; **17**(15-16):1767-1777.
16. Lin Z. An issue of statistical analysis in controlled multi-centre studies: How shall we weight the centres? *Statistics in Medicine* 1999; **18**(4):365-373.
17. Schaubel DE. Variance estimation for clustered recurrent event data with a small number of clusters. *Statistics in Medicine* 2005; **24**(19):3037-3051.
18. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**(15-16):1753-1765.
19. Yamaguchi T, Ohashi Y, Matsuyama Y. Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Statistical Methods in Medical Research* 2002; **11**(3):221-236.
20. Rosenthal D, Frank JD. Psychotherapy and the placebo effect. *Psychological Bulletin* 1956; **53**:294-302.
21. Thorne FC. Rules of evidence in the evaluation of the effect of psychotherapy. *Journal of Clinical Psychology* 1952; **8**:3841.
22. Horvath P. Placebos and common factors in two decades of psychotherapy research. *Psychological Bulletin* 1988; **104**(2):214-225.

23. Castonguay LG, Holtforth MG. Change in psychotherapy: A plea for no more "nonspecific" and false dichotomies. *Clinical Psychology-Science & Practice* 2005; **12**(2):198-201.
24. Craighead WE, Sheets ES, Bjornsson AS, Arnarson EO. Specificity and nonspecificity in psychotherapy. *Clinical Psychology-Science & Practice* 2005; **12**(2):189-193.
25. DeRubeis RJ, Brotman MA, Gibbons CJ. A conceptual and methodological analysis of the nonspecifics argument. *Clinical Psychology-Science & Practice* 2005; **12**(2):174-183.
26. Kazdin AE. Treatment outcomes, common factors, and continued neglect of mechanisms of change. *Clinical Psychology-Science & Practice* 2005; **12**(2):184-188.
27. Wampold BE. Establishing specificity in psychotherapy scientifically: Design and evidence issues. *Clinical Psychology-Science & Practice* 2005; **12**(2):194-197.
28. Luborsky L, Singer B, Luborsky L. Comparative studies of psychotherapies: Is it true that "everyone has won and all must have prizes"? *Archives of General Psychiatry* 1975; **32**:995-1008.
29. Luborsky L. Are common factors across different psychotherapies the main explanation for the Dodo Bird verdict that "Everyone has won so all shall have prizes"? *Clinical Psychology: Science and Practice* 1995; **2**(1):-109.
30. Wampold BE, Ahn H-N, Coleman HLK. Medical model as metaphor: Old habits die hard. *Journal of Counseling Psychology* 2001; **48**(3):268-273.
31. Wampold BE, Minami T, Tierney SC, Baskin TW, Bhati KS. The placebo is powerful: Estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *Journal of Clinical Psychology* 2005; **61**(7):835-854.
32. Wampold BE, Mondin GW, Moody M, Stich F, Benson K, Ahn H-N. A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes.". *Psychological Bulletin* 1997; **122**(3):203-215.
33. Wampold BE. Contextualizing psychotherapy as a healing practice: Culture, history, and methods. *Applied & Preventive Psychology* 2001; **10**(2):69-86.
34. Wampold BE. Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research* 1997; **7**(1):21-43.
35. Wampold BE. Root metaphor versus square root: Research evidence for a contextualist theme. *Journal of Counseling & Development* 1991; **70**(2):297-299.
36. Wampold BE, Mondin GW, Moody M, Ahn H-N. The flat earth as a metaphor for the evidence for uniform efficacy of bona fide psychotherapies: Reply to Crits-Christoph (1997) and Howard et al. (1997). *Psychological Bulletin* 1997; **122**(3):226-230.
37. Bower P, Rowland N. Effectiveness and cost effectiveness of counselling in primary care. *Cochrane Database of Systematic Reviews* 2006; Issue 3. Art. No.: CD001025. DOI: 10.1002/14651858.CD001025.pub2.
38. Blitstein JL, Murray DM, Hannan PJ, Shadish WR. Increasing the degrees of freedom in future group randomized trials: The df* approach. *Evaluation Review* 2005; **29**(3):268-286.
39. Kwong, G. P. S and Higgins, J. P. T. Adjusting for clustering in meta-analysis of individually-randomized trials. 2008.
Ref Type: Unpublished Work
40. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis* 2006; **50**:3681-3701.
41. Huynh, C. L. A unified approach to the estimation of effect size in meta-analysis. Annual Meeting of the American Educational Research Association. 1989.
Ref Type: Conference Proceeding
42. White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clinical Trials* 2005; **2**(2):141-151.
43. Hedges LV. Effect sizes in cluster randomized designs. *Journal of Educational and Behavioral Statistics* 2007; **32**(4):341-370.
44. Goldstein H, Yang M, Omar R, Turner R, Thompson S. Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics* 2000; **49**(3):399-412.

45. Kiesler DJ. Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin* 1966; **65**(2):110-136.
46. Meehl PE. Psychotherapy. *Annual Review of Psychology* 1955; **6**:357-378.
47. Paul GL, Licht MH. Resurrection of uniformity assumption myths and the fallacy of statistical absolutes in psychotherapy research. *Journal of Consulting & Clinical Psychology* 1978; **46**(6):1531-1534.
48. Elkin I, Parloff MB, Hadley SW, Autry JH. NIMH treatment of depression collaborative research program. Background and Research Plan. *Archives of General Psychiatry* 1985; **42**(3):305-316.
49. Serlin RC, Wampold BE, Levin JR. Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: a comment on Siemer and Joormann (2003). *Psychological Methods* 2003; **8**(4):524-534.
50. Luborsky L, McLellan AT, Diguer L, Woody GE, Seligman DA. The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology-Science & Practice* 1997; **4**(1):53-65.
51. Luborsky L. The personality of the psychotherapist. *Menninger Quarterly.* 1952; **6**:1-6.
52. Parloff M, Waskow I, Wolfe B. Research on therapist variables in relation to process and outcome. In: Garfield S. L., Bergin AE (eds) *Handbook of psychotherapy and behavior change.* Wiley: New York, 1978; pp 233-282.
53. Beutler LE, Malik M, Alimohamed S, Harwood TM, Talebi H, Noble S, Wong E. Therapist Variables. In: Lambert M. J. (ed) *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change.* Wiley & Sons: New York, 2004; pp 227-306.
54. Ricks DF. Supershrink: Methods of a therapist judged successful on the basis of adult outcome of adolescent patients. In: Ricks D. F., Roff M, Thomas A (eds) *Life history research in psychopathology.* University of Minnesota Press: Minneapolis, 1974.
55. Howard KI, Orlinsky DE, Perilstein J. Contribution of therapists to patients' experiences in psychotherapy: A components of variance model for analyzing process data. *Journal of Consulting & Clinical Psychology* 1976; **44**(4):520-526.
56. Orlinsky DE, Howard KI. Gender and psychotherapeutic outcome. In: Brodsky A. M., Hare-Muslin RT (eds) *Women and psychotherapy.* Guilford: New York, 1980; pp 3-34.
57. Brooker C, Wiggins RD. Nurse therapist trainee variability: the implications for selection and training. *Journal of Advanced Nursing* 1983; **8**:321-328.
58. Luborsky L, Crits-Christoph P, McLellan T, Woody G, Piper W, Imber S, Liberman B. Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry* 1986; **56**(4):501-512.
59. Luborsky L, McLellan AT, Woody GE, O'Brien CP, Auerbach A. Therapist success and its determinants. *Archives of General Psychiatry* 1985; **42**(6):602-611.
60. McLellan AT, Woody G, Luborsky L, Goehl L. Is the counsellor an "active ingredient" in methadone treatment? An examination of treatment success among four counselors. *Journal of Nervous and Mental Disease.* 1988; **176**:423-430.
61. Shapiro DA, Firth-Cozens J, Stiles WB. The question of therapists' differential effectiveness. A Sheffield psychotherapy project addendum. *British Journal of Psychiatry* 1989; **154**(MAR):383-385.
62. Okiishi J, Lambert MJ, Nielsen SL, Ogles BM. Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy* 2003; **10**(6):361-373.
63. Okiishi JC, Lambert MJ, Eggett D, Nielsen L, Dayton DD, Vermeersch DA. An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology* 2006; **62**(9):1157-1172.
64. Wampold BE, Brown GS. Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting & Clinical Psychology* 2005; **73**(5):914-923.
65. McKay KM, Imel ZE, Wampold BE. Psychiatrist effects in the psychopharmacological treatment of depression. *Journal of Affective Disorders* 2006; **92**(2-3):287-290.
66. Lutz W, Leon SC, Martinovich Z, Lyons JS, Stiles WB. Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology* 2007; **54**(1):32-39.

67. Baldwin SA, Wampold BE, Imel ZE. Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting & Clinical Psychology* 2007; **75**(6):842-852.
68. Dinger U, Strack M, Leichsenring F, Wilmers F, Schauenburg H. Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology* 2008; **64**(3):344-354.
69. Crits-Christoph P, Gallop R. Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research* 2006; **16**(2):178-181.
70. Elkin I, Falconnier L, Martinovich Z, Mahoney C. Rejoinder to commentaries by Stephen Soldz and Paul Crits-Christoph on therapist effects. *Psychotherapy Research* 2006; **16**(2):182-183.
71. Elkin I, Falconnier L, Martinovich Z, Mahoney C. Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research* 2006; **16**(2):144-160.
72. Kim D-M, Wampold BE, Bolt DM. Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research* 2006; **16**(2):161-172.
73. Soldz S. Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research* 2006; **16**(2):173-177.
74. Wampold BE, Bolt DM. Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research* 2006; **16**(2):184-187.
75. Elkin I, Falconnier L, Martinovich Z. Misrepresentation in Wampold and Bolt's critique of Elkin, Falconnier, Martinovich, and Mahoney's study of therapist effects. *Psychotherapy Research* 2007; **17**(2):253-256.
76. Wampold BE, Bolt DM. Appropriate estimation of therapist effects: One more time. *Psychotherapy Research* 2007; **17**(2):256-257.
77. Wampold BE, Bolt DM. The consequences of "anchoring" in longitudinal multilevel models: Bias in the estimation of patient variability and therapist effects. *Psychotherapy Research* 2007; **17**(5):509-514.
78. Elkin I. A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. *Clinical Psychology-Science & Practice* 1999; **6**(1):10-32.
79. Martindale C. The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting & Clinical Psychology* 1978; **46**(6):1526-1530.
80. Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting & Clinical Psychology* 1991; **59**(1):20-26.
81. Roberts C. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Statistics in Medicine* 1999; **18**(19):2605-2615.
82. Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. *British Medical Journal* 2005; **330**(7483):142-144.
83. Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials* 2005; **2**(2):163-173.
84. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials* 2005; **2**(2):152-162.
85. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, for the CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacolgic treatment: Explanation and elaboration. *Annals of Internal Medicine* 2008; **148**:295-309.
86. MRC Health Services and Public Health Board. A Framework for Development and Evaluation of RCTs for Complex Interventions to Improve Health. 2000.
Ref Type: Report
87. Schnurr PP, Friedman MJ, Engel CC, Foa EB, Shea MT, Chow BK, Resick PA, Thurston V, Orsillo SM, Haug R, Turner C, Bernardy N. Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *JAMA* 2007; **297**(8):820-830.

88. Kubany ES, Hill EE, Owens JA. Cognitive trauma therapy for battered women with PTSD. *Journal of Consulting & Clinical Psychology* 2004; **72**(1):3-18.
89. Cohen JA, Mannarino AP. A treatment outcome study for sexually abused preschool children: initial findings. *Journal of the American Academy of Child & Adolescent Psychiatry* 1996; **35**(1):42-50.
90. Lovell K, Cox D, Haddock G, Jones C, Raines D, Garvey R, Roberts C, Hadley S. Telephone administered cognitive behaviour therapy for treatment of obsessive compulsive disorder: Randomised controlled non-inferiority trial. *British Medical Journal* 2006; **333**(7574):883-886.
91. Ekbohm G, Melander H. The subject-by-formulation interaction as a criterion of interchangeability of drugs. *Biometrics* 1989; **45**(4):1249-1254.
92. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health* 2004; **94**(3):416-422.
93. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Statistics in Medicine* 2008; **27**(27):5578-5585.
94. Turner RM, White IR, Croudace T, for the PIP Study Group. Analysis of cluster randomized cross-over trial data: A comparison of methods. *Statistics in Medicine* 2007; **26**:274-289.
95. Goldstein H. *Multilevel statistical models*. Arnold: London, 2003.
96. Wampold BE, Serlin RC. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods* 2000; **5**(4):425-433.
97. Crits-Christoph P, Tu X, Gallop R. Therapists as fixed versus random effects-some statistical and conceptual issues: A comment on Siemer and Joormann (2003). *Psychological Methods* 2003; **8**(4):518-523.
98. Kirk RE. *Experimental Design Procedures for the Behavioral Sciences*. Brooks/Cole: Belmont, California, 1968.
99. Winer BJ. *Statistical Principles in Experimental Design*. McGraw-Hill: New York, 1971.
100. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* 1996; **49**(4):435-439.
101. Venning P, Durie A, Roland M, Roberts C, Leese B. Randomized controlled trial comparing cost effectiveness of general practitioners and nurse practitioners in primary care. *British Medical Journal* 2000; **320**:1048-1053.
102. Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. Wiley: New York, 2002.
103. Dunn OJ, Clark VA. *Applied Statistics: Analysis of Variance and Regression*. John Wiley & Sons: New York, 1987.
104. Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques. *Statistics in Medicine* 1997; **16**(18):2063-2079.
105. Cook JA, Ramsay CR, Fayers P. Statistical evaluation of learning curve effects in surgical trials. *Clinical Trials* 2004; **1**:421-427.
106. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325-337.
107. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000.
108. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 1946; **2**(6):110-114.
109. Hoover DR. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. *Statistics in Medicine* 2002; **21**(10):1351-1364.
110. DiSantostefano RL, Muller KE. A comparison of power approximations for Satterthwaite's test. *Communications in Statistics - Simulation* 1995; **24**(3):583-593.
111. Elashof, J. D. nQuery Advisor Version 5.0 User's Guide. 2002. Los Angeles, CA.
Ref Type: Report
112. Moser BK, Stevens GR, Watts CL. The two-sample t test versus Satterthwaite's approximate F-test. *Communications in Statistics - Theory and Methods* 1989; **18**:3963-3975.

113. http://personalpages.manchester.ac.uk/staff/Chris.Roberts/ . 23-4-2008. Ref Type: Electronic Citation

114. Moerbeek M, Wong WK. Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine* 2008; **27**(15):2850-2864.

115. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: Implications for sample size calculations. *Statistics in Medicine* 2001; **20**(3):377-390.

116. Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* 2001; **43**(1):75-86.

117. Elkin I, Pilkonis PA, Docherty JP, Sotsky SM. Conceptual and methodological issues in comparative studies of psychotherapy and pharmacotherapy, I: Active ingredients and mechanisms of change. *American Journal of Psychiatry* 1988; **145**(8):909-917.

118. Elkin I, Pilkonis PA, Docherty JP, Sotsky SM. Conceptual and methodological issues in comparative studies of psychotherapy and pharmacotherapy, II: Nature and timing of treatment effects. *American Journal of Psychiatry* 1988; **145**(9):1070-1076.

119. Elkin I, Shea MT, Watkins JT, Imber SD, Sotsky SM, Collins JF, Glass DR, Pilkonis PA, Leber WR, Docherty JP, Fiester SJ, Parloff MB. National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Archives of General Psychiatry* 1989; **46**(11):971-982.

120. Blanchard EB, Hickling EJ, Devineni T, Veazey CH, Galovski TE, Mundy E, Malta LS, Buckley TC. A controlled evaluation of cognitive behaviorial therapy for posttraumatic stress in motor vehicle accident survivors. *Behaviour Research & Therapy* 2003; **41**(1):79-96.

121. Browne W, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. *Statistical Modelling.* 2001; **1**:103-124.

122. Goldstein H, Rasbash J, Browne W, Woodhouse G, Poulain M. Multilevel models in the study of dynamic household structures. *European Journal of Population* 2001; **16**:373-387.

123. Rasbash J, Steele F, Browne W, Prosser B. *A User's Guide to MLwiN Version 2.0*. Institute of Education: London, 2004.

124. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [Updated September 2006]. In: John Wiley & Sons, Ltd: Chichester, UK, 2006.

125. Siemer M, Joormann J. Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods* 2003; **8**(4):497-517.

126. Staines GL. Comparative outcome evaluations of psychotherapies: Guidelines for addressing eight limitations of the gold standard of causal inference. *Psychotherapy: Theory, Research, Practice, Training* 2007; **44**(2):161-174.

127. UKATT Research Team. United Kingdom Alcohol Treatment Trial (UKATT): hypotheses, design and methods. *Alcohol & Alcoholism* 2001; **36**(1):11-21.

128. Schnurr PP, Friedman MJ, Lavori PW, Hsieh FY. Design of Department of Veterans Affairs Cooperative Study No. 420: Group treatment of posttraumatic stress disorder. *Controlled Clinical Trials* 2001; **22**(1):74-88.

129. Schnurr PP, Friedman MJ, Engel CC, Foa EB, Shea MT, Resick PM, James KE, Chow BK. Issues in the design of multisite clinical trials of psychotherapy: VA Cooperative Study No. 494 as an example. *Contemporary Clinical Trials* 2005; **26**(6):626-636.

130. Lambert MJ. The individual therapist's contribution to psychotherapy process and outcome. *Clinical Psychology Review* 1989; **9**(4):469-485.

131. Blowers C, Cobb J, Mathews A. Generalised anxiety: A controlled treatment study. *Behaviour Research & Therapy* 1987; **25**(6):493-502.

132. Borkovec TD, Mathews AM, Chambers A, Ebrahimi S, Lytle R, Nelson R. The effects of relaxation training with cognitive or nondirective therapy and the role of relaxation-induced anxiety in the treatment of generalized anxiety. *Journal of Consulting & Clinical Psychology* 1987; **55**(6):883-888.

133. Durham RC, Turvey AA. Cognitive therapy vs behaviour therapy in the treatment of chronic general anxiety. *Behaviour Research & Therapy* 1987; **25**(3):229-234.

134. Butler G, Fennell M, Robson P, Gelder M. Comparison of behavior therapy and cognitive behavior therapy in the treatment of generalized anxiety disorder. *Journal of Consulting & Clinical Psychology* 1991; **59**(1):167-175.
135. Barlow DH, Rapee RM, Brown TA. Behavioral treatment of generalized anxiety disorder. *Behavior Therapy* 1992; **23**:551-570.
136. Durham RC, Murphy T, Allan T, Richard K, Treliving LR, Fenton GW. Cognitive therapy, analytic psychotherapy and anxiety management training for generalised anxiety disorder. *British Journal of Psychiatry* 1994; **165**(3):315-323.
137. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *British Medical Journal* 2001; **322**:355-357.
138. Brien CJ, Bailey RA. Multiple randomizations. *Journal of the Royal Statistical Society, Series B* 2006; **68**(4):571-609.
139. Kazdin AE. Comparative outcome studies of psychotherapy: Methodological issues and strategies. *Journal of Consulting & Clinical Psychology* 1986; **54**(1):95-105.
140. Wilkins W. Therapy-therapist confounds in psychotherapy research. *Cognitive Therapy and Research* 1986; **10**(1):3-11.
141. Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, Yusuf S, Sackett DL, Cina CS, Walter SD, Haynes B, Schunemann HJ, Norman GR, Guyatt GH. Need for expertise based randomised controlled trials. *British Medical Journal* 2005; **330**(7482):88-91.
142. Siemer M, Joormann J. Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: Reply to Crits-Christoph, Tu, and Gallop (2003) and Serlin, Wampold, and Levin (2003). *Psychological Methods* 2003; **8**(4):535-544.
143. Miettinen OS. The clinical trial as a paradigm for epidemiologic research. *Journal of Clinical Epidemiology* 1989; **42**:491-496.
144. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M., Davey-Smith G, Altman DG (eds) *Systematic reviews in health care: Meta-analysis in context.* BMJ Books: London, 2001; pp 285-312.
145. Glass GV. Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; **5**(10):3-8.
146. Birge RT. The calculation of errors by the method of least squares. *Phys.Rev.* 1932; **16**:1-32.
147. Cochran WG. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society* 1937; **4 (Supplement)**:102-118.
148. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation: New York, 2009.
149. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods* 2002; **7**(1):105-125.
150. Morris, S. B. Effect size estimation from pretest-posttest-control designs with heterogeneous variances. 20th Annual Conference of the Society for Industrial and Organizational Psychology. 2005.
Ref Type: Conference Proceeding
151. Morris SB. Estimating effect sizes from pretrest-posttest-control group designs. *Organizational Research Methods* 2008; **11**(2):364-386.
152. Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials* 2009; **6**:16-27.
153. Curtin F, Altman DG, Elbourne D. Meta-analysis combining parallel and cross-over clinical trials. I: Continuous outcomes. *Statistics in Medicine* 2002; **21**(15):2131-2144.
154. Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. III: The issue of carry-over. *Statistics in Medicine* 2002; **21**(15):2161-2173.
155. Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. II: Binary outcomes. *Statistics in Medicine* 2002; **21**(15):2145-2159.

156. Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: Methodological issues. *International Journal of Epidemiology* 2002; **31**(1):140-149.
157. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Statistical Methods in Medical Research* 2001; **10**(5):325-338.
158. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Statistics in Medicine* 2002; **21**(19):2971-2980.
159. Hedges LV. Effect sizes in nested designs. In: Cooper Harris, Hedges LV, Valentine JC (eds) *The Handbook of Research Synthesis and Meta-Analysis.* Russell Sage Foundation: New York, 2009; pp 337-355.
160. Zou G. One relative risk versus two odds ratios: implications for meta-analyses involving paired and unpaired binary data. *Clinical Trials* 2007; **4**(1):25-31.
161. Laopaiboon M. Meta-analyses involving cluster randomization trials: A review of published literature in health care. *Statistical Methods in Medical Research* 2003; **12**(6):515-530.
162. Huppert FA, Van Niekerk JK, Herbert J. Dehydroepiandrosterone (DHEA) supplementation for cognition and well-being. *Cochrane Database of Systematic Reviews* 2001; (1).
163. The Vitamin A and Pneumonia Working Group. Potential interventions for the prevention of childhood pneumonia in developing countries: a meta-analysis of data from field trials to assess the impact of vitamin A supplementation on pneumonia morbidity and mortality. *Bulletin of the World Health Organization* 1995; **73**(5):609-619.
164. Fawzi WW, Chalmers T, Herrera MG, Mosteller F. Vitamin A supplementation and child mortality: a meta-analysis. *The Journal of the American Medical Association.* 1993; **269**(7):898-903.
165. Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50: a quality assessment and meta-analysis. *The Medical Journal of Australia.* 1995; **162**:625-629.
166. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics.* 1992; **48**:577-585.
167. Hendrick R, Smith R, Rutledge IJ, Smart C. Benefit of screening mammography in women aged 40-49: A new meta-analysis of randomized controlled trials. *Journal of the National Cancer Institute Monographs* 1997; **22**:87-92.
168. Kramer M. Balanced protein/energy supplementation in pregnancy. *The Cochrane Library, Issue 1* 1999.
169. Higgins, J. P. T. and Green, S. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]. http://www.cochrane.org/resources/handbook/hbook.htm . 2005. 31-5-2005.
Ref Type: Electronic Citation
170. Starr M, Chalmers I, Clarke M, Oxman AD. The origins, evolution, and future of The Cochrane Database of Systematic Reviews. *International Journal of Technology Assessment in Health Care* 2009; **25**(Supplement 1):182-195.
171. Bacaltchuk J, Hay P, Trefiglio R. Antidepressants versus psychological treatments and their combination for bulimia nervosa. *Cochrane Database of Systematic Reviews* 2001; Issue 4. Art. No.: CD003385. DOI: 10.1002/14651858.CD003385.
172. Furukawa TA, Watanabe N, Churchill R. Combined psychotherapy plus antidepressants for panic disorder with or without agoraphobia. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD004364. DOI: 10.1002/14651858.CD004364.pub2.
173. Morriss RK, Faizal MA, Jones AP, Williamson PR, Bolton C, McCarthy JP. Interventions for helping people recognise early signs of recurrence in bipolar disorder. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD004854. DOI: 10.1002/14651858.CD004854.pub2.
174. Glasscoe CA, Quittner AL. Psychological interventions for cystic fibrosis. *Cochrane Database of Systematic Reviews* 2003; Issue 3. Art. No.: CD003148. DOI: 10.1002/14651858.CD003148.

175. Edwards AGK, Hailey S, Maxwell M. Psychological interventions for women with metastatic breast cancer. *Cochrane Database of Systematic Reviews* 2004; Issue 2. Art. No.: CD004253. DOI: 10.1002/14651858.CD004253.pub2.
176. Rees K, Bennett P, West R, Davey SG, Ebrahim S. Psychological interventions for coronary heart disease. *Cochrane Database of Systematic Reviews* 2004; Issue 2. Art. No.: CD002902. DOI: 10.1002/14651858.CD002902.pub2.
177. Kisely S, Campbell LA, Skerritt P. Psychological interventions for symptomatic management of non-specific chest pain in patients with normal coronary anatomy. *Cochrane Database of Systematic Reviews* 2005; Issue 1. Art. No.: CD004101. DOI: 10.1002/14651858.CD004101.pub2.
178. Shaw K, O'Rourke P, Del MC, Kenardy J. Psychological interventions for overweight or obesity. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD003818. DOI: 10.1002/14651858.CD003818.pub2.
179. Yorke J, Fleming S, Shuldham C. Psychological interventions for children with asthma. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD003272. DOI: 10.1002/14651858.CD003272.pub2.
180. Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis. *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD004431. DOI: 10.1002/14651858.CD004431.pub2.
181. Uman LS, Chambers CT, McGrath PJ, Kisely S. Psychological interventions for needle-related procedural pain and distress in children and adolescents. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD005179. DOI: 10.1002/14651858.CD005179.pub2.
182. Yorke J, Fleming SL, Shuldham CM. Psychological interventions for adults with asthma. *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD002982. DOI: 10.1002/14651858.CD002982.pub3.
183. Eccleston C, Yorke L, Morley S, Williams AC, Mastroyannopoulou K. Psychological therapies for the management of chronic and recurrent pain in children and adolescents. *Cochrane Database of Systematic Reviews* 2003; Issue 1. Art. No.: CD003968. DOI: 10.1002/14651858.CD003968.
184. Binks CA, Fenton M, McCarthy L, Lee T, Adams CE, Duggan C. Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD005652. DOI: 10.1002/14651858.CD005652.
185. Hunot V, Churchill R, Silva-de LM, Teixeira V. Psychological therapies for generalised anxiety disorder. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD001848. DOI: 10.1002/14651858.CD001848.pub4.
186. Bisson J, Andrew M. Psychological treatment of post-traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD003388. DOI: 10.1002/14651858.CD003388.pub2.
187. Amato L, Minozzi S, Davoli M, Vecchi S, Ferri M, Mayet S. Psychosocial and pharmacological treatments versus pharmacological treatments for opioid detoxification. *Cochrane Database of Systematic Reviews* 2004; Issue 4. Art. No.: CD005031. DOI: 10.1002/14651858.CD005031.
188. Dennis CL, Creedy D. Psychosocial and psychological interventions for preventing postpartum depression. *Cochrane Database of Systematic Reviews* 2004; Issue 4. Art. No.: CD001134. DOI: 10.1002/14651858.CD001134.pub2.
189. Amato L, Minozzi S, Davoli M, Vecchi S, Ferri M, Mayet S. Psychosocial combined with agonist maintenance treatments versus agonist maintenance treatments alone for treatment of opioid dependence. *Cochrane Database of Systematic Reviews* 2004; Issue 4. Art. No.: CD004147. DOI: 10.1002/14651858.CD004147.pub2.
190. Jeffery DP, Ley A, McLaren S, Siegfried N. Psychosocial treatment programmes for people with both severe mental illness and substance misuse. *Cochrane Database of Systematic Reviews* 2000; Issue 2. Art. No.: CD001088. DOI: 10.1002/14651858.CD001088.
191. Hay PJ, Bacaltchuk J, Stefano S. Psychotherapy for bulimia nervosa and binging. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD000562. DOI: 10.1002/14651858.CD000562.pub2.

192. Hajek P, Stead LF. Aversive smoking for smoking cessation. *Cochrane Database of Systematic Reviews* 2001; Issue 3. Art.No.: CD000546. DOI: 10.1002/14651858.CD000546.pub2.

193. Brazzelli M, Griffiths P. Behavioural and cognitive interventions with or without other treatments for the management of faecal incontinence in children. *Cochrane Database of Systematic Reviews* 2006; Issue 2. Art. No.: CD002240. DOI: 10.1002/14651858.CD002240.pub3.

194. O'Kearney RT, Anstey KJ, von SC. Behavioural and cognitive behavioural therapy for obsessive compulsive disorder in children and adolescents. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD004856. DOI: 10.1002/14651858.CD004856.pub2.

195. Ostelo RWJG, van-Tulder MW, Vlaeyen JWS, Linton SJ, Morley SJ, Assendelft WJJ. Behavioural treatment for chronic low-back pain. *Cochrane Database of Systematic Reviews* 2005; Issue 1. Art. No.: CD002014. DOI: 10.1002/14651858.CD002014.pub2.

196. Macdonald GM, Higgins JPT, Ramchandani P. Cognitive-behavioural interventions for children who have been sexually abused. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD001930. DOI: 10.1002/14651858.CD001930.pub2.

197. Jones C, Cormac I, Silveira-da-Mota-Neto-JI, Campbell C. Cognitive behaviour therapy for schizophrenia. *Cochrane Database of Systematic Reviews* 2004; Issue 4. Art. No.: CD000524. DOI: 10.1002/14651858.CD000524.pub2.

198. Montgomery P, Dennis J. Cognitive behavioural interventions for sleep problems in adults aged 60+. *Cochrane Database of Systematic Reviews* 2003; Issue 1. Art. No.: CD003161. DOI: 10.1002/14651858.CD003161.

199. James A, Soler A, Weatherall R. Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD004690. DOI: 10.1002/14651858.CD004690.pub2.

200. Martinez DP, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD005233. DOI: 10.1002/14651858.CD005233.pub2.

201. Joy CB, Adams CE, Rice K. Crisis intervention for people with severe mental illnesses. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD001087. DOI: 10.1002/14651858.CD001087.pub3.

202. Baldwin C, Parsons T, Logan S. Dietary advice for illness-related malnutrition in adults. *Cochrane Database of Systematic Reviews* 2001; Issue 1. Art. No.: CD002008. DOI: 10.1002/14651858.CD002008.pub2.

203. Brunner EJ, Thorogood M, Rees K, Hewitt G. Dietary advice for reducing cardiovascular risk. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD002128. DOI: 10.1002/14651858.CD002128.pub2.

204. Crawford-Walker CJ, King A, Chan S. Distraction techniques for schizophrenia. *Cochrane Database of Systematic Reviews* 2005; Issue 1. Art. No.: CD004717. DOI: 10.1002/14651858.CD004717.pub2.

205. Woolfenden SR, Williams K, Peat J. Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17. *Cochrane Database of Systematic Reviews* 2001; Issue 2. Art.No.:CD003015.DOI: 10.1002/14651858.CD003015.

206. Pharoah F, Mari J, Rathbone J, Wong W. Family intervention for schizophrenia. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD000088. DOI: 10.1002/14651858.CD000088.pub2.

207. Stead LF, Lancaster T. Group behaviour therapy programmes for smoking cessation. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD001007. DOI: 10.1002/14651858.CD001007.pub2.

208. Izquierdo-de SA, Khan M. Hypnosis for schizophrenia. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD004160. DOI: 10.1002/14651858.CD004160.pub2.

209. Abbot NC, Stead LF, White AR, Barnes J. Hypnotherapy for smoking cessation. *Cochrane Database of Systematic Reviews* 1998; Issue 2. Art. No.: CD001008. DOI: 10.1002/14651858.CD001008.

210. Lancaster T, Stead LF. Individual behavioural counselling for smoking cessation. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD001292. DOI: 10.1002/14651858.CD001292.pub2.

211. Barbato A, D'Avanzo B. Marital therapy for depression. *Cochrane Database of Systematic Reviews* 2006; Issue 2. Art. No.: CD004188. DOI: 10.1002/14651858.CD004188.pub2.

212. He Y, Li C. Morita therapy for schizophrenia. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD006346. DOI: 10.1002/14651858.CD006346.

213. Littell JH, Popa M, Forsythe B. Multisystemic Therapy for social, emotional, and behavioral problems in youth aged 10-17. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD004797. DOI: 10.1002/14651858.CD004797.pub4.

214. Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder. *Cochrane Database of Systematic Reviews* 2006; Issue 2. Art. No.: CD004381. DOI: 10.1002/14651858.CD004381.pub2.

215. Gold C, Heldal TO, Dahle T, Wigram T. Music therapy for schizophrenia or schizophrenia-like illnesses. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD004025. DOI: 10.1002/14651858.CD004025.pub2.

216. Barlow J, Coren E, Stewart-Brown SSB. Parent-training programmes for improving maternal psychosocial health. *Cochrane Database of Systematic Reviews* 2003; Issue 4. Art. No.: CD002020. DOI: 10.1002/14651858.CD002020.pub2.

217. Eustice S, Roe B, Paterson J. Prompted voiding for the management of urinary incontinence in adults. *Cochrane Database of Systematic Reviews* 2000; Issue 2. Art. No.: CD002113. DOI: 10.1002/14651858.CD002113.

218. Rose S, Bisson J, Churchill R, Wessely S. Psychological debriefing for preventing post traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews* 2002; Issue 2. Art. No.: CD000560. DOI: 10.1002/14651858.CD000560.

219. Spector A, Orrell M, Davies S, Woods B. Reality orientation for dementia. *Cochrane Database of Systematic Reviews* 2000; Issue 3. Art. No.: CD001119. DOI: 10.1002/14651858.CD001119.

220. Woods B, Spector A, Jones C, Orrell M, Davies S. Reminiscence therapy for dementia. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD001120. DOI: 10.1002/14651858.CD001120.pub2.

221. Abbass AA, Hancock JT, Henderson J, Kisely S. Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD004687. DOI: 10.1002/14651858.CD004687.pub3.

222. Buckley LA, Pettit T. Supportive therapy for schizophrenia. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art.No.: CD004716. DOI: 10.1002/14651858.CD004716.pub2.

223. Stead LF, Perera R, Lancaster T. Telephone counselling for smoking cessation. *Cochrane Database of Systematic Reviews* 2006; Issue 3. Art. No.: CD002850. DOI: 10.1002/14651858.CD002850.pub2.

224. Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol. *Cochrane Database of Systematic Reviews* 2003; Issue 3. Art. No.: CD001366. DOI: 10.1002/14651858.CD001366.

225. Rice VH, Stead LF. Nursing interventions for smoking cessation. *Cochrane Database of Systematic Reviews* 2004; Issue 1. Art. No.: CD001188. DOI: 10.1002/14651858.CD001188.pub2.

226. den-Boer PCAM, Wiersma D, Russo S, van-den-Bosch RJ. Paraprofessionals for anxiety and depressive disorders. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD004688. DOI: 10.1002/14651858.CD004688.pub2.

227. Carr AB, Ebbert JO. Interventions for tobacco cessation in the dental setting. *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD005084. DOI: 10.1002/14651858.CD005084.pub2.

228. Huibers MJH, Beurskens AJHM, Bleijenberg G, Schayck Cv. Psychosocial interventions delivered by general practitioners. *Cochrane Database of Systematic Reviews* 2003; Issue 2. Art. No.: CD003494. DOI: 10.1002/14651858.CD003494.

229. Ray KL, Hodnett ED. Caregiver support for postpartum depression. *Cochrane Database of Systematic Reviews* 2001; Issue 2. Art. No.: CD000946. DOI: 10.1002/14651858.CD000946.

230. Smith CA, Collins CT, Cyna AM, Crowther CA. Complementary and alternative therapies for pain management in labour. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD003521. DOI: 10.1002/14651858.CD003521.pub2.

231. Glazener CMA, Evans JHC, Cheuk DKL. Complementary and miscellaneous interventions for nocturnal enuresis in children. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD005230. DOI: 10.1002/14651858.CD005230.

232. Glazener CMA, Evans JHC, Peto RE. Complex behavioural and educational interventions for nocturnal enuresis in children. *Cochrane Database of Systematic Reviews* 2004; Issue 1. Art. No.: CD004668. DOI: 10.1002/14651858.CD004668.

233. Hunter KF, Moore KN, Cody DJ, Glazener CMA. Conservative management for postprostatectomy urinary incontinence. *Cochrane Database of Systematic Reviews* 2004; Issue 2. Art. No.: CD001843. DOI: 10.1002/14651858.CD001843.pub2.

234. Perry A, Coulton S, Glanville J, Godfrey C, Lunn J, McDougall C, Neale Z. Interventions for drug-using offenders in the courts, secure establishments and the community. *Cochrane Database of Systematic Reviews* 2006; Issue 3. Art. No.: CD005193. DOI: 10.1002/14651858.CD005193.pub2.

235. Oakley-Browne MA, Adams P, Mobberley PM. Interventions for pathological gambling. *Cochrane Database of Systematic Reviews* 2000; Issue 1. Art. No.: CD001521. DOI: 10.1002/14651858.CD001521.

236. Anderson CS, Hackett ML, House AO. Interventions for preventing depression after stroke. *Cochrane Database of Systematic Reviews* 2004; Issue 2. Art. No.: CD003689. DOI: 10.1002/14651858.CD003689.pub2.

237. Dinh ZT, Goss C, Heitman E, Roberts I, DiGuiseppi C. Interventions for preventing injuries in problem drinkers. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD001857. DOI: 10.1002/14651858.CD001857.pub2.

238. Ebbert JO, Rowland LC, Montori V, Vickers KS, Erwin PC, Dale LC, Stead LF. Interventions for smokeless tobacco use cessation. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD004306. DOI: 10.1002/14651858.CD004306.pub2.

239. Hackett ML, Anderson CS, House AO. Interventions for treating depression after stroke. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD003437. DOI: 10.1002/14651858.CD003437.pub2.

240. Forbes C, Jepson R, Martin HP. Interventions targeted at women to encourage the uptake of cervical screening. *Cochrane Database of Systematic Reviews* 1999; Issue 3. Art. No.: CD002834. DOI: 10.1002/14651858.CD002834.

241. Johnson WD, Hedges LV, Diaz RM. Interventions to modify sexual risk behaviors for preventing HIV infection in men who have sex with men. *Cochrane Database of Systematic Reviews* 2002; Issue 4. Art. No.: CD001230. DOI: 10.1002/14651858.CD001230.

242. Faulkner G, Cohn T, Remington G. Interventions to reduce weight gain in schizophrenia. *Cochrane Database of Systematic Reviews* 2007; Issue 1. Art. No.: CD005148. DOI: 10.1002/14651858.CD005148.pub2.

243. Norris SL, Zhang X, Avenell A, Gregg E, Schmid CH, Lau J. Long-term non-pharmacological weight loss interventions for adults with prediabetes. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD005270. DOI: 10.1002/14651858.CD005270.

244. Marine A, Ruotsalainen J, Serra C, Verbeek J. Preventing occupational stress in healthcare workers. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD002892. DOI: 10.1002/14651858.CD002892.pub2.

245. Merry S, McDowell H, Hetrick S, Bir J, Muller N. Psychological and/or educational interventions for the prevention of depression in children and adolescents. *Cochrane Database of Systematic Reviews* 2004; Issue 2. Art. No.: CD003380. DOI: 10.1002/14651858.CD003380.pub2.

246. Hawton K, Townsend E, Arensman E, Gunnell D, Hazell P, House A, Van HK. Psychosocial and pharmacological treatments for deliberate self harm. *Cochrane*

Database of Systematic Reviews 1999; Issue 4.Art.No.:CD001764.DOI: 10.1002/14651858.CD001764.

247. Hajek P, Stead LF, West R, Jarvis M, Lancaster T. Relapse prevention interventions for smoking cessation. *Cochrane Database of Systematic Reviews* 2005; Issue 1. Art. No.: CD003999. DOI: 10.1002/14651858.CD003999.pub2.
248. Perkins SJ, Murphy R, Schmidt U, Williams C. Self-help and guided self-help for eating disorders. *Cochrane Database of Systematic Reviews* 2006; Issue 3. Art. No.: CD004191. DOI: 10.1002/14651858.CD004191.pub2.
249. Hodnett ED, Fredericks S. Support during pregnancy for women at increased risk of low birthweight babies. *Cochrane Database of Systematic Reviews* 2003; Issue 3. Art. No.: CD000198. DOI: 10.1002/14651858.CD000198.
250. Lima MS, Reisser-Lima AAP, Soares BGO, Farrell M. Antidepressants for cocaine dependence. *Cochrane Database of Systematic Reviews* 2003; Issue 2. Art. No.: CD002950. DOI: 10.1002/14651858.CD002950.
251. Bize R, Burnand B, Mueller Y, Cornuz J. Biomedical risk assessment as an aid for smoking cessation. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD004705. DOI: 10.1002/14651858.CD004705.pub2.
252. Lima RA, Lima MS, Soares BGO, Farrell M. Carbamazepine for cocaine dependence. *Cochrane Database of Systematic Reviews* 2002; Issue 2. Art. No.: CD002023. DOI: 10.1002/14651858.CD002023.
253. Gourlay SG, Stead LF, Benowitz NL. Clonidine for smoking cessation. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD000058. DOI: 10.1002/14651858.CD000058.pub2.
254. O'Connor AM, Stacey D, Entwistle V, Llewellyn TH, Rovner D, Holmes RM, Tait V, Tetroe J, Fiset V, Barry M, Jones J. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews* 2003; Issue 1. Art. No.: CD001431. DOI: 10.1002/14651858.CD001431.
255. Soares BGO, Lima MS, Lima RA, Farrell M. Dopamine agonists for cocaine dependence. *Cochrane Database of Systematic Reviews* 2003; Issue 2. Art. No.: CD003352. DOI: 10.1002/14651858.CD003352.
256. Jolliffe JA, Rees K, Taylor RS, Thompson D, Oldridge N, Ebrahim S. Exercise-based rehabilitation for coronary heart disease. *Cochrane Database of Systematic Reviews* 2001; Issue 1. Art. No.: CD001800. DOI: 10.1002/14651858.CD001800.
257. Larun L, Nordheim LV, Ekeland E, Hagen KB, Heian F. Exercise in prevention and treatment of anxiety and depression among children and young people. *Cochrane Database of Systematic Reviews* 2006; Issue 3. Art. No.: CD004691. DOI: 10.1002/14651858.CD004691.pub2.
258. Ekeland E, Heian F, Hagen KB, Abbott J, Nordheim L. Exercise to improve self-esteem in children and young people. *Cochrane Database of Systematic Reviews* 2004; Issue 1. Art. No.: CD003683. DOI: 10.1002/14651858.CD003683.pub2.
259. Doggett C, Burrett S, Osborn DA. Home visits during pregnancy and after birth for women with an alcohol or drug problem. *Cochrane Database of Systematic Reviews* 2005; Issue 4. Art. No.: CD004456. DOI: 10.1002/14651858.CD004456.pub2.
260. Ebrahim S, Beswick A, Burke M, Davey SG. Multiple risk factor interventions for primary prevention of coronary heart disease. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD001561. DOI: 10.1002/14651858.CD001561.pub2.
261. Silagy C, Lancaster T, Stead L, Mant D, Fowler G. Nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews* 2004; Issue 3. Art. No.: CD000146. DOI: 10.1002/14651858.CD000146.pub2.
262. Srisurapanont M, Jarusuraisin N. Opioid antagonists for alcohol dependence. *Cochrane Database of Systematic Reviews* 2005; Issue 1. Art. No.: CD001867. DOI: 10.1002/14651858.CD001867.pub2.
263. Minozzi S, Amato L, Vecchi S, Davoli M, Kirchmayer U, Verster A. Oral naltrexone maintenance treatment for opioid dependence. *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD001333. DOI: 10.1002/14651858.CD001333.pub2.

264. Riemsma RP, Kirwan JR, Taal E, Rasker JJ. Patient education for adults with rheumatoid arthritis. *Cochrane Database of Systematic Reviews* 2003; Issue 2. Art. No.: CD003688. DOI: 10.1002/14651858.CD003688.
265. Stein DJ, Ipser JC, Seedat S. Pharmacotherapy for post traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews* 2006; Issue 1. Art. No.: CD002795. DOI: 10.1002/14651858.CD002795.pub2.
266. O'Brien K, Nixon S, Glazier RH, Tynan AM. Progressive resistive exercise interventions for adults living with HIV/AIDS. *Cochrane Database of Systematic Reviews* 2004; Issue 4. Art. No.: CD004248. DOI: 10.1002/14651858.CD004248.pub2.
267. Lacasse Y, Goldstein R, Lasserson TJ, Martin S. Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews* 2006; Issue 4. Art. No.: CD003793. DOI: 10.1002/14651858.CD003793.pub2.
268. Lancaster T, Stead LF. Self-help interventions for smoking cessation. *Cochrane Database of Systematic Reviews* 2005; Issue 3. Art. No.: CD001118. DOI: 10.1002/14651858.CD001118.pub2.
269. Glazener CMA, Evans JHC, Peto R. Tricyclic and related drugs for nocturnal enuresis in children. *Cochrane Database of Systematic Reviews* 2003; Issue 3. Art. No.: CD002117. DOI: 10.1002/14651858.CD002117.
270. Crowther R, Marshall M, Bond G, Huxley P. Vocational rehabilitation for people with severe mental illness. *Cochrane Database of Systematic Reviews* 2001; Issue 2. Art. No.: CD003080. DOI: 10.1002/14651858.CD003080.
271. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment* 1999; **3**(5):iii-92.
272. Mulrow CD, Oxman AD. *Cochrane Collaboration Handbook*. Update Software: Oxford, 1996.
273. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996; **17**(1):1-12.
274. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**(5):408-412.
275. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal* 2001; **323**(7303):42-46.
276. Van Tulder MW, Assendelft WJ, Koes BW, Bouter LM, et al. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders. *Spine* 1997; **22**:2323-2330.
277. Van Tulder MW, Furlan A, Bombardier C, Bouter L, et al. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine* 2003; **28**(12):1290-1299.
278. Moncrieff J, Churchill R, Drummond C, McGuire H. Development of a quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research.* 2001; **10**:126-133.
279. Grant AM, Cody DJ, Glazener CMA, Hay-Smith J, Herbison P, Lapitan MC, et al. Cochrane Incontinence Group. In: *The Cochrane Library, Issue 1.* Update Software: Oxford, 2003.
280. Centre for Reviews and Dissemination. *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for Carrying Out or Commissioning Reviews (CRD Report Number 4)*. 1996.
281. Centre for Reviews and Dissemination. *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for Carrying Out or Commissioning Reviews (CRD Report Number 4)*. 2001.
282. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality of both randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health* 1998; **52**(6):377-384.

283. Kenardy J, Carr V. Imbalance in the debriefing debate: What we don't know far outweights what we do. *Bulletin of the Australian Psychological Society* 1996; **17**(1):4-6.
284. Churchill R, Wessely S, Lewis G. Pharmacotherapy and psychotherapy for depression. In: *The Cochrane Library, 4.* Update Software: Oxford, 1997.
285. Edwards A, Hood K, Matthews EJ, Russell D, Russell IT, Barker J, et al. The effectiveness of one-to-one risk communication interventions in health care: A systematic review. *Medical Decision Making* 2000; **20**:290-297.
286. Walwyn R, Roberts C. Therapist variation in randomised trials of psychotherapy: Implications for precision, internal and external validity. *Statistical Methods in Medical Research* 2010; **19**(3):291-315.
287. Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry* 2000; **176**:249-252.
288. Moher D, Liberati A, Tetzlaff J, Altman DGatPG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine* 2009; **151**(4):264-269.
289. Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., Thase, M. E., Perkins, P., and Leon, A. C. A new scale for assessing the quality of randomized clinical trials of psychotherapy. Comprehensive Psychiatry . 2009.
Ref Type: In Press
290. London School of Economics Centre for Economic Performance Mental Health Policy Group. The Depression Report: A new deal for depression and anxiety disorders. http://cep.lse.ac.uk/ . 2005.
Ref Type: Electronic Citation
291. Bird A. *We Need to Talk: The case for psychological therapy on the NHS*. Mental Health Foundation: London, 2006.
292. CSIP Choice and Access Team. *Improving Access to Psychological Therapies: Positive Practice Guide*. Department of Health: London, 2007.
293. Department of Health. *Improving Access to Psychological Therapies Implementation Plan*. Department of Health: London, 2008.
294. Wooster E. *While We Are Waiting: Experiences of waiting for and receiving psychological therapies on the NHS*. Mental Health Foundation: London, 2008.
295. Pringle M, Laverty J. A counsellor in every practice? Reasons for caution. *British Medical Journal* 1993; **306**:2-3.
296. Bond T. The nature and role of counselling in primary care. In: Keithley J., Bond T, Marsh G (eds) *Counselling in Primary Care.* Oxford University Press: Oxford, 2002; pp 3-24.
297. Wessely S. The rise of counselling and the return of alienism. *BMJ* 1996; **313**(7050):158-160.
298. Mellor-Clark J, Simms-Ellis R, Burton M. *National Survey of Counsellors in Primary Care: Evidence for growing professionalisation?* Royal College of General Practitioners: London, 2001.
299. Department of Health. *Treatment Choice in Psychological Therapies and Counselling: Evidence based clinical practice guideline*. Department of Health: London, 2001.
300. Counselling in Primary Care Trust. *Work Specification for Counsellors Working in GP Practices*. Counselling in Primary Care Trust: Staines, 1992.
301. NHS Centre for Reviews and Dissemination. Counselling in primary care. *Effectiveness Matters* 2001; **5**(2).
302. British Association for Counselling. *Guidelines for the Employment of Counsellors in General Practice*. BAC: Rugby, 1993.
303. The Counselling in Primary Care Trust. *Referral Guidelines for Counselling in General Practice: Supplement 1*. CPCT: Staines, 1995.
304. The Scottish Office Department of Health National Medical Advisory Committee. *A Report by the National Medical Advisory Committee*. Department of Health: Edinburgh, 1998.

305. McLeod J. *The Work of Counsellors in General Practice [Occasional Paper 37]*. Royal College of General Practitioners: London, 1988.
306. Naji SA, Atherton-Naji A, Beattie JAG, Donald PM. Counselling in Scottish General Practice: A national sample survey. *Primary Care Psychiatry* 1998; **4**(3):133-139.
307. Sibbald B, Addington-Hall J, Brenneman D, Freeling P. *The Role of Counsellors Working in General Practice [Occasional Paper 74]*. Royal College of General Practitioners: London, 1996.
308. Clark A, Hook J, Stein K. Counsellors in primary care in Southampton: A questionnaire survey of their qualifications, working arrangements and casemix. *British Journal of General Practice* 1997; **47**:613-617.
309. Sibbald B, Addington-Hall J, Brenneman D, Freeling P. Counsellors in English and Welsh practices: Their nature and distribution. *British Medical Journal* 1993; **306**:29-33.
310. Rowland N. Counselling and counselling skills. In: Sheldon M. (ed) *Counselling in General Practice.* Royal College of General Practitioners: London, 1992; pp 1-7.
311. Smith S, Norton K. *Counselling Skills for Doctors*. Open University Press: Buckingham, UK, 1999.
312. Chambless DL, Hollon SD. Defining empirically supported therapies. *Journal of Consulting & Clinical Psychology* 1998; **66**(1):7-18.
313. Barrowclough C, King P, Colville J, Russell E, Burns A, Tarrier N. A randomized trial of the effectiveness of cognitive-behavioral therapy and supportive counseling for anxiety symptoms in older adults. *Journal of Consulting & Clinical Psychology* 2001; **69**(5):756-762.
314. Boot D, Gillies P, Fenelon J, Reubin R, Wilkins M, Gray P. Evaluation of the short-term impact of counseling in general practice. *Patient Education & Counseling* 1994; **24**(1):79-89.
315. Chilvers C, Dewey M, Fielding K, Gretton V, Miller P, Palmer B, Weller D, Churchill R, Williams I, Bedi N, Duggan C, Lee A, Harrison G. Antidepressant drugs and generic counselling for treatment of major depression in primary care: Randomised trial with patient preference arms. *British Medical Journal* 2001; **322**(7289):772-775.
316. Friedli K, King MB, Lloyd M, Horder J. Randomised controlled assessment of non-directive psychotherapy versus routine general-practitioner care. *Lancet* 1997; **350**(9092):1662-1665.
317. Harvey I, Nelson SJ, Lyons RA, Unwin C, Monaghan S, Peters TJ. A randomized controlled trial and economic evaluation of counselling in primary care. *British Journal of General Practice* 1998; **48**(428):1043-1048.
318. Hemmings A. Counselling in primary care: A randomised controlled trial. *Patient Education & Counseling* 1997; **32**(3):219-230.
319. King M, Sibbald B, Ward E, Bower P, Lloyd M, Gabbay M, Byford S. Randomised controlled trial of non-directive counselling, cognitive-behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care. *Health Technology Assessment* 2000; **4**(19):1-83.
320. Simpson S, Corney R, Fitzgerald P, Beecham J. A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression. *Health Technology Assessment* 2000; **4**(36).
321. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11):1539-1558.
322. Brewin T, Bradley C. Patient preferences and randomised clinical trials. *British Medical Journal* 1989; **299**:313-315.
323. Beck AT, Ward C, Mendelson M, Erbaugh J. An inventory for measuring depression. *Archives of General Psychiatry* 1961; **6**:561-571.
324. Blenkiron P. Does the management of depression in general practice match current guidelines? *Primary Care Psychiatry* 1998; **4**:121-125.
325. Donoghue J, Tylee A. The treatment of depression: Prescribing patterns of antidepresants in primary care in the UK. *British Journal of Psychiatry* 1996; **168**:164-168.
326. Scott J, Moon CA, Blacker CV, Thomas JM, Scott AIF, Freeman CPL. Edinburgh Primary Care Depression Study. *British Journal of Psychiatry* 1994; **164**:410-415.

327. Sibbald, B., Addington-Hall, J., Brenneman, D., and Freeling, P. The role of counsellors in general practice. Occasional Paper No.: 74. 1996. London, Royal College of General Practitioners.
Ref Type: Report

328. NHS Executive. NHS Psychotherapy Services in England: A review of strategic policy. 1996. London, NHS Executive.
Ref Type: Report

329. Roth A, Parry G. The implications of psychotherapy research for clinical practice and service development: Lessons and limitations. *Journal of Mental Health* 1997; **6**:367-380.

330. Rice N, Leyland A. Multilevel models: Applications to health data. *Journal of Health Services & Research Policy* 1996; **1**:154-164.

331. Campbell M, Grimshaw J. Cluster randomised trials: Time for improvement. *British Medical Journal* 1998; **317**:1171.

332. Wood J, Freemantle N. Choosing an appropriate unit of analysis in trials of interventions that attempt to influence practice. *Journal of Health Services & Research Policy* 1999; **4**:44-48.

333. Roth A, Fonagy P. *What works for whom?* The Guildford Press: New York, 1996.

334. Kline P. Problems of methodology in studies of psychotherapy. In: Dryden W., Feltham C (eds) *Psychotherapy and Its Discontents.* Open University Press: Bristol, 1992.

335. Bower P, Byford S, Barber J, Beecham J, Simpson S, Friedli K, Corney R, King M, Harvey I. Meta-analysis of data on costs from trials of counselling in primary care: Using individual patient data to overcome sample size limitations in economic analyses. *British Medical Journal* 2003; **326**(7401):1247-1250.

336. Roberts, C., Roberts, S. A., and Vail, A. Clustering effects in an individually randomised trial: Design considerations for the "therapist effect". 2006.
Ref Type: Conference Proceeding

337. Cornfield J. Randomization by group: A formal analysis. *American Journal of Epidemiology* 1978; **108**(2):100-102.

338. Jarman B, Hurtwitz B, Cook A, Bajekal M, Lee A. Effects of community based nurses specialising in Parkinson's disease on health outcome and costs: randomised controlled trial. *BMJ* 2002; **324**:1072-1075.

339. Bliese PD, Halverson RR. Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management* 1998; **24**(2):157-172.

340. Cohen J. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement* 1973; **33**:107-112.

341. Kennedy JJ. The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement* 1970; **30**:855-889.

342. Keren G, Lewis C. Partial omega squared for ANOVA designs. *Educational and Psychological Measurement* 1979; **39**:119-128.

343. Olejnik S, Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 2003; **8**(4):434-447.

344. Pearson K. On a correction needful in the case of the correlation ratio. *Biometrika* 1911; **8**:254-256.

345. Goodyer I, Dubicka B, Wilkinson P, Kelvin R, Roberts C, Byford S, Breen S, Ford C, Barrett B, Leech A, Rothwell J, White L, Harrington R. Selective serotonin reuptake inhibitors (SSRIs) and routine specialist care with and without cognitive behaviour therapy in adolescents with major depression: randomised controlled trial. *BMJ* 2007; **335**:142-149.

346. Strupp HH. Success and failure in time-limited psychotherapy: A systematic comparison of two cases - comparison 1. *Archives of General Psychiatry* 1980; **37**:595-603.

347. Strupp HH. Success and failure in time-limited psychotherapy. A systematic comparison of two cases - comparison 2. *Archives of General Psychiatry* 1980; **37**:708-716.

348. Strupp HH. Success and failure in time-limited psychotherapy: With special reference to a lay counselor. *Archives of General Psychiatry* 1980; **37**:831-841.

349. Strupp HH. Success and failure in time-limited psychotherapy: A systematic comparison of two cases - comparison 4. *Archives of General Psychiatry* 1980; **37**:947-954.

350. Blatt SJ, Sanislow III CA, Zuroff DC, Pilkonis PA. Characteristics of effective therapists: Further analyses of data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting & Clinical Psychology* 1996; **64**(6):1276-1284.

351. Project MATCH Research Group. Therapist effects in three treatments for alcohol problems. *Psychotherapy Research* 1998; **8**(4):455-474.

352. Huppert JD, Bufka LF, Barlow DH, Gorman JM, Shear MK, Woods SW. Therapists, therapist variables, and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting & Clinical Psychology* 2001; **69**(5):747-755.

353. Garfield SL. The therapist as a neglected variable in psychotherapy research. *Clinical Psychology: Science and Practice* 1997; **4**(1):40-43.

354. Bergin AE. Neglect of the therapist and the human dimensions of change: A commentary. *Clinical Psychology: Science and Practice* 1997; **4**(1):83-89.

355. Strupp HH, Anderson T. On the limitations of therapy manuals. *Clinical Psychology: Science and Practice* 1997; **4**(1):76-82.

356. Lambert MJ, Okiishi JC. The effects of the individual psychotherapist and implications for future research. *Clinical Psychology: Science and Practice* 1997; **4**(1):66-75.

357. Beutler LE. The psychotherapist as a neglected variable in psychotherapy: An illustration by reference to the role of therapist experience & training. *Clinical Psychology: Science and Practice* 1997; **4**(1):44-52.

358. Okiishi, J. C. Waiting for supershrink: An empirical analysis of therapist effects. Dissertation Abstracts International: Section B: The Sciences and Engineering Vol 61(9-B), 4999. 2001.
Ref Type: Thesis/Dissertation

359. Schoenwald SK, Letourneau EJ, Halliday-Boykins C. Predicting therapist adherence to a transported family-based treatment for youth. *Journal of Clinical Child and Adolescent Psychology* 2005; **34**(4):658-670.

360. Stiles, W. B., Barkham, M., Mellor-Clark, J., Connell, J., and Lutz, W. Therapists' differential effectiveness in routine practice in United Kingdom National Health Service settings. American Psychology Association Convention. 2007.
Ref Type: Conference Proceeding

361. Kim, D.-M. Therapist effects and treatment effects in psychotherapy: Analysis on the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP). 2003.
Ref Type: Thesis/Dissertation

362. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health* 2004; **94**(3):393-399.

363. Crits-Christoph P, Baranackie K, Kurcias JS, Beck AT, Carroll KM, Perry K, Luborsky L, McLellan AT, Woody GE, Thompson L, Gallagher D, Zitrin C. Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research* 1991; **1**(2):81-91.

364. Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J., Abramowitz, J. S., Andersson, G., Atkins, D. C., Carlbring, P., Carroll, K. M., Christensen, A., Eddington, K. M., Ehlers, A., Feaster, D. J., Keijsers, G. P. J., Koch, E., Kuyken, W., Lange, A., Lincoln, T., Stephens, R. S., Taylor, S., Trepka, C., and Watson, J. Estimates of intraclass correlations associated with therapists: An initial database to guide power calculations. 2009.
Ref Type: Unpublished Work

365. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology* 2004; **57**(8):785-794.

366. Campbell MK, Grimshaw JM, Steen N, for the Changing Professional Practice in Europe Group. Sample size calculations for cluster randomised trials. *Journal of Health Services Research and Policy* 2000; **5**:12-16.

367. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: The case of implementation research. *Clinical Trials* 2005; **2**(2):99-107.
368. Donner A. An empirical study of cluster randomization. *International Journal of Epidemiology* 1982; **11**:283-286.
369. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlation for the design of community-based surveys and intervention studies. *American Journal of Epidemiology* 1999; **149**:876-883.
370. Hedges LV, Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis* 2007; **29**(1):60-87.
371. Murray DM, Rooney BL, Hannan PJ, Peterson AV, Ary DV, Biglan A, Botvin GJ, Evans RI, Flay BR, Futterman R, Getz JG, Marek PM, Orlandi M, Pentz MA, Perry CL, Schinke SP. Intraclass correlation among common measures of adolescent smoking: Estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology* 1994; **140**(11):1038-1050.
372. Murray DM, Blistein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review* 2003; **27**(1):79-103.
373. Murray DM, Catellier DJ, Hannan PJ, Treuth MS, Stevens J, Schmitz KH, Rice JC, Conway TL. School-level intraclass correlation for physical activity in adolescent girls. *Medicine and Science in Sports and Exercise* 2004; **36**(5):876-882.
374. Murray DM, Stevens J, Hannan PJ, Catellier DJ, Schmitz KH, Dowda M, Conway TL, Rice JC, Yang S. School-level intraclass correlation for physical activity in sixth grade girls. *Medicine and Science in Sports and Exercise* 2006; **38**(5):926-936.
375. Nye B, Konstantopoulos S, Hedges LV. How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis* 2004; **26**(3):237-257.
376. Verma V, Le T. An analysis of sampling errors for the demographic and health surveys. *International Statistical Review* 1996; **64**:265-294.
377. Fisher RA. *Statistical Methods for Research Workers*. Oliver & Boyd: Edinburgh, 1954.
378. Donner A, Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* 1986; **42**(2):401-412.
379. Weinberg R, Patel YC. Simulated intraclass correlation coefficients and their z-transforms. *Journal of Statistical Computation and Simulation* 1981; **13**(1):13-26.
380. DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177-188.
381. Fisher RA. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1921; **1**:3-32.
382. Rosner B, Donner A, Hennekens CH. Estimation of interclass correlation from familial data. *Applied Statistics* 1977; **26**:179-187.
383. Wang CS, Yandell BS, Rutledge JJ. Bias of maximum likelihood estimator of intraclass correlation. *Theoretical and Applied Genetics* 1991; **82**:421-424.
384. Ponzoni RW, James JW. Possible biases in heritability estimates from intraclass correlation. *Theoretical and Applied Genetics* 1978; **53**:25-27.
385. Ginsburg EH. On the planning of the experiment on estimation of intraclass correlation. *Biom.Z.* 1973; **15**:47-52.
386. Wang CS, Yandell BS, Rutledge JJ. The dilemma of negative analysis of variance estimators of intraclass correlation. *Theoretical and Applied Genetics* 1992; **85**:79-88.
387. Konishi S. Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics* 1985; **37**:87-94.
388. Konishi S, Gupta AK. Testing the equality of several intraclass correlation coefficients. *Journal of Statistical Planning and Inference* 1989; **21**:93-105.
389. Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. *Biometrics* 1980; **36**(1):19-25.
390. Kempthorne O, Tandon OB. The estimation of heritability by regression of offspring on parent. *Biometrics* 1953; **9**:90-100.
391. Donner A, Koval JJ. The large sample variance of an intraclass correlation. *Biometrika* 1980; **67**:719-722.

392. Gill JL, Jensen EL. Probability of obtaining negative estimates of heritability. *Biometrics* 1968; **24**:517-526.
393. Searle SR. Topics in variance component estimation. *Biometrics* 1971; **27**:1-76.
394. Murray DM, Hannan PJ, Baker WL. A Monte Carlo study of alternative responses to intraclass correlation in community trials. Is it ever possible to avoid Cornfield's penalties? *Evaluation Review* 1996; **20**:313-337.
395. Konishi S, Khatri CG, Rao CR. Inferences on multivariate measures of interclass and intraclass correlations in familial data. *Journal of the Royal Statistical Society, Series B* 1991; **53**:649-659.
396. Donner A, Zou G. Testing the equality of dependent intraclass correlation correlation coefficients. *The Statistician* 2002; **51**(3):367-379.
397. Snedecor GW, Cochran WG. *Statistical Methods*. Iowa University Press: Ames, Iowa, 1989.
398. Thomas JD, Hultquist RA. Interval estimation for the unbalanced case of the one-way random effects model. *Annals of Statistics* 1978; **6**:582-587.
399. Bhargava RP. Tests of significance for intraclass correlation when family sizes are not equal. *Sankhya* 1946; **7**:435-438.
400. Spjotvoll E. Optimum invariant tests in unbalanced variance component models. *Annals of Mathematical Statistics* 1967; **38**(2):422-428.
401. Donner A, Wells GA, Eliasziw M. On two approximations to the F-distribution: Application to testing for intraclass correlation in family studies. *The Canadian Journal of Statistics* 1989; **17**(2):209-215.
402. Abramowitz JS, Foa EB, Franklin ME. Exposure and ritual prevention for obsessive-compulsive disorder: Effects of intensive versus twice-weekly sessions. *Journal of Consulting & Clinical Psychology* 2003; **71**:394-398.
403. Carlbring P, Bohman S, Brunt S, Buhrman M, Westling BE, Ekselius L, et al. Remote treatment of panic disorder: A randomized trial of internet-based cognitive behavior therapy supplemented with telephone calls. *American Journal of Psychiatry* 2006; **163**(12):2119-2125.
404. Ehlers A, Clark DM, Hackmann A, McManus F, Fennell M, Herbert C, et al. A randomized controlled trial of cognitive therapy, a self-help booklet, and repeated assessments as early interventions for posttraumatic stress disorder. *Archives of General Psychiatry* 2003; **60**:1024-1032.
405. Marijuana Treatment Project Research Group. Brief treatments for cannabis dependence: Findings from a randomized multisite trial. *Journal of Consulting & Clinical Psychology* 2004; **72**:455-466.
406. Taylor S, Thordarson DS, Maxfield L, Fedoroff IC, Lovell K, Ogrodniczuk JS. Comparative efficacy, speed, and adverse effects of three PTSD treatments: Exposure therapy, EMDR, and relaxation training. *Journal of Consulting & Clinical Psychology* 2003; **71**:330-338.
407. van Minnen A, Hoogduin KA, Keijsers GP, Hellenbrand I, Hendriks GJ. Treatment of trichotillomania with behavioral therapy or fluoxetine: A randomized, waiting-list controlled study. *Archives of General Psychiatry* 2003; **60**:517-522.
408. Watson JC, Gordon LB, Stermac L, Kalogerakos F, Steckley P. Comparing the effectiveness of process-experiential with cognitive-behavioral psychotherapy in the treatment of depression. *Journal of Consulting & Clinical Psychology* 2003; **71**:773-781.
409. Kuyken W. Cognitive therapy outcome: The effects of hopelessness in a naturalistic outcome study. *Behaviour Research & Therapy* 2004; **42**:631-646.
410. Lincoln TM, Rief W, Hahlweg K, Frank M, von Witzleben I, Schroeder B, et al. Effectiveness of an empirically supported treatment for social phobia in the field. *Behaviour Research & Therapy* 2003; **41**:1251-1269.
411. Merrill KA, Tolbert VE, Wade WA. Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting & Clinical Psychology* 2003; **71**:404-409.

412. Trepka C, Rees A, Shapiro DA, Hardy GE, Barkham M. Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy & Research* 2004; **28**:143-157.
413. Lange A, Rietdijk D, Hudcovicova M, van de Ven JP, Schrieken B, Emmelkamp PM. Interapy: A controlled randomized trial of the standardized treatment of posttraumatic stress through the internet. *Journal of Consulting & Clinical Psychology* 2003; **71**:901-909.
414. Szapocznik J, Feaster DJ, Mitrani VB, Prado G, Smith L, Robinson-Batista C, et al. Structural ecosystems therapy for HIV-seropositive African American women: Effects on psychological distress, family hassles, and family support. *Journal of Consulting & Clinical Psychology* 2004; **72**:288-303.
415. Hedges LV. Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics* 2007; **32**(2):151-179.
416. Hedges LV. Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics* 2007; **32**(2):151-179.
417. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: Extension to cluster randomised trials. *British Medical Journal* 2004; **328**(7441):702-708.
418. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomized trials: For discussion. *Statistics in Medicine* 2001; **20**(3):489-496.
419. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999; **28**(2):319-326.
420. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006; **35**:1292-1300.
421. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619-629.
422. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; **30**(3):261-293.
423. Horn SD, Horn RA, Duncan DB. Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association* 1975; **70**:380-385.
424. Royall RM, Cumberland WG. Variance estimation in finite population sampling. *Journal of the American Statistical Association* 1978; **73**:351-358.
425. Borenstein M. Effect sizes for continuous data. In: Cooper H., Hedges LV, Valentine JC (eds) *The Handbook of Research Synthesis and Meta-analysis.* Russell Sage Foundation: New York, 2009; pp 221-236.
426. Kim S-H, Cohen AS. On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics* 1998; **23**(4):356-377.
427. Rosner B. A generalization of the paired t-test. *Applied Statistics* 1982; **31**:9-13.
428. Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine* 2001; **20**(15):2219-2241.
429. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine* 1999; **18**(20):2693-2708.
430. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**(11):1559-1573.
431. Bohning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics* 2002; **3**(4):445-457.
432. Hedges LV. Estimation of effect size from a series of independent experiments. *Psychological Bulletin* 1982; **92**(2):490-499.
433. Glass GV. Integrating findings: The meta-analysis of research. In: Peacock F. E. (ed) *Review of Research in Education.* Itasca, Ill, 1977.
434. McGaw B, Glass GV. Choice of the metric for effect size in meta-analysis. *American Educational Research Journal* 1980; **17**(3):325-337.
435. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press: New York, 1977.

436. Grissom RJ, Kim JJ. Review of assumptions and problems in appropriate conceptualization of effect size. *Psychological Methods* 2001; **6**:135-146.
437. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 1981; **6**(2):107-128.
438. Becker BJ. Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology* 1988; **41**:257-278.
439. Carlson KD, Schmidt FL. Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology* 1999; **84**:851-862.
440. Marshall M, Crowther R, Almaraz-Serrano A, Creed F, Sledge W, Kluiter H, Roberts C, Hill E, Wiersma D, Bond GR, Huxley P, Tyrer P. Systematic reviews of the effectiveness of day care for people with severe mental disorders: (1) acute day hospital versus admission; (2) vocational rehabilitation; (3) day hospital versus outpatient care. *Health Technology Assessment* 2001; **5**(21):1-75.
441. Birks J, Flicker L. Selegiline for Alzheimer's disease. *Cochrane Database of Systematic Reviews* 2003; Issue 1. Art. No.: CD000442.
442. Early Supported Discharge Trialists. Services for reducing duration of hospital care for acute stroke patients. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD000443.
443. Hedges LV. A random effects model for effect sizes. *Psychological Bulletin* 1983; **93**(2):388-395.
444. Johnson NL, Welch BL. Applications of the noncentral t-distribution. *Biometrika* 1939; **31**:362-389.
445. Box GEP. Some theorems on quadratic forms applied to the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* 1954; **25**:290-302.
446. Hogben D, Pinkman RS, Wilk MB. The moments of the non-central t-distribution. *Biometrika* 1961; **48**(3/4):465-468.
447. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology* 1986; **123**(2):203-208.
448. Edwards AL, Cronbach LJ. Experimental design for research in psychotherapy. *Journal of Clinical Psychology* 1952; **8**(1):51-59.
449. Thorne FC. Rules of evidence in the evaluation of the effects of psychotherapy. *Journal of Clinical Psychology* 1952; **8**(1):38-41.
450. Watson RI. Research design and methodology in evaluating the results of psychotherapy. *Journal of Clinical Psychology* 1952; **8**(1):29-33.
451. Watson RI. Measuring the effectiveness of psychotherapy: Problems for investigation. *Journal of Clinical Psychology* 1952; **8**(1):60-64.
452. Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregate data. *Psychological Methods* 2009; **14**(2):165-176.
453. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal* 2010; **340**:c221.
454. Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials* 2005; **2**(3):209-217.
455. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions* 2002; **25**(1):76-97.
456. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2000; **19**(24):3417-3432.
457. Egger M, Davey-Smith G, Altman DG. *Systematic Reviews in Health Care: Meta-analysis in context*. BMJ Books: London, 2001.
458. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Process and product optimization using designed experiments*. Wiley: New York, 2009.
459. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**:277-303.

460. Morris SB. Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology* 2000; **53**:17-29.