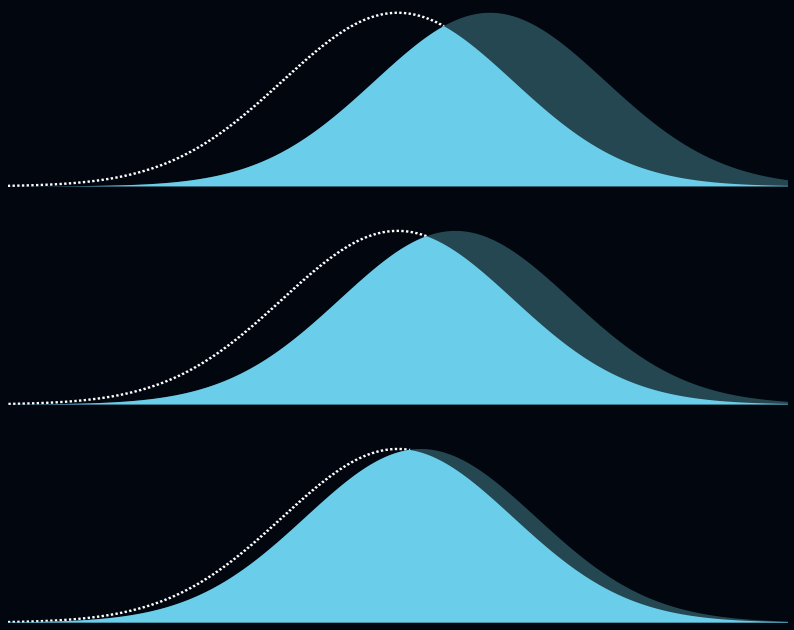
Thesis for doctoral degree (Ph.D.)
2019

Methodological Issues in Psychological Treatment Research

Applications to Gambling Research and Therapist Effects



Kristoffer Magnusson

Karolinska Institutet

Thesis for doctoral degree (Ph.D.) 2019
Methodological Issues in Psychological Treatment Research
Kristoffer Magnusson

From DEPARTMENT OF CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

METHODOLOGICAL ISSUES IN PSYCHOLOGICAL TREATMENT RESEARCH

APPLICATIONS TO GAMBLING RESEARCH AND THERAPIST EFFECTS

Kristoffer Magnusson

Stockholm 2019

© 2019 Kristoffer Magnusson
All previously published papers were reproduced with permission from the publisher
Published by Karolinska Institutet
Written in R Markdown using the Bookdown package
Printed by US-AB
ISBN 978-91-7831-604-5

Methodological issues in psychological treatment research: Applications to gambling research and therapist effects
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Kristoffer Magnusson

Principal Supervisor:
Prof. Per Carlbring
Stockholm University
Department of Psychology

Co-supervisors:
Prof. Clara Hellner
Karolinska Institutet
Department of Clinical Neuroscience

Prof. Gerhard Andersson
Linköping University
Department of Behavioral Sciences and Learning

Opponent:
Prof. Andy Field
University of Sussex
School of Psychology

Examination Board:
Assoc. Prof. Lene Lindberg
Karolinska Institutet
Department of Public Health Sciences

Assoc. Prof. Cecilia Åslund
Uppsala University
Centre for Clinical Research

Prof. Anders C Håkansson
Lund University
Faculty of Medicine

Abstract

Over the last couple of decades evidence-based psychotherapies have flourished, and there are now therapies that are well-established for a wide range of problems. At the same time the mental-health burden is still enormous, and challenges to the dissemination of treatments are substantial. Despite the considerable gains in knowledge that have been made, many issues remain unsolved, and there are many reasons to be skeptical of the current quality of the evidence.

The aim of this thesis was to explore methodological challenges that impact the evaluation of psychological treatments in general but also gambling treatment trials specifically. In the first part, I present a broader overview of some of the contemporary issues that concern the scientific investigation of psychotherapies. Two issues are then explored in more detail, 1) the broader issue of therapist effects in longitudinal studies, and 2) the more specific issue of analyzing semicontinuous data as a treatment outcome. After expanding on these issues, the findings are then used in two clinical gambling studies.

In **Study I**, we investigated the consequences of ignoring therapist effects in longitudinal data. We derived what factors impact the type I errors, and performed an extensive simulation study. The empirical simulation results validated the analytical results and showed that even when 5% of the variance in slopes is at the therapist level, the type I errors can be substantially inflated. When analyzing data from longitudinal studies, investigators should account for the possibility that therapists might have different overall slopes. In an LMM, this can be accounted for by including a random slope at the therapist level. In order to help investigators plan multilevel longitudinal studies, an R package (`powerlmm`) was developed.

In **Study II**, we investigated the challenges of estimating treatment effects in gambling studies using gambling expenditure as an outcome. Gambling outcomes are typically very skewed and can include a large number of zeros. Investigators typically try to analyze such data mostly by log transforming the outcome, or continue with a standard analysis based on normally distributed residuals. In this paper, we propose that a marginalized two-part model can be a more attractive option. We compared the performance of the proposed two-part model to the typical methods used by investigators. The performance of these models were compared using real data and via different Monte Carlo simulation scenarios. The choice of an appropriate estimand for treatment effects was also discussed, and we argue that gambling researchers should primarily be concerned with the overall reduction in gambling losses.

In **Study III**, we applied and extended the work in Study II to investigate how concordant gamblers and their concerned significant others (CSOs) were in their reports of gambling losses. The sample consisted of problem gamblers and their CSOs participating in a trial comparing individual CBT versus behavioral couples therapy. A total of 133 dyads were included, and we used their baseline reports of gambling losses using the timeline followback covering the last 30 days. Overall we found that there was a fair level of agreement, ICC = .57, 95% CI [.48, .64]. There were some evidence that partner CSOs had a higher level of agreement compared to parent CSOs, ICC_{diff} = .20, 95% CI [.03, .39].

In **Study IV**, we applied the results from Study I, II, and III to investigate the effects

of an internet-delivered program aimed at the CSOs of treatment refusing problem gamblers. In total, 100 CSOs of treatment-refusing problem gamblers were randomized to either ten weeks of ICBT or a waitlist control. At posttest the intervention group reported an improvement on the CSO's emotional consequences (d = -0.90, 95% CI [-1.47, -0.33]), relationship satisfaction (d = 0.41, 95% [0.05, 0.76]), anxiety (d = -0.45, 95% [-0.81, -0.09]), depression (d = -0.49, 95% [-0.82, -0.16]). Any effects on the CSO's reports on gambling losses and on treatment-seeking were inconclusive. Problem gamblers are hard to influence via their CSO proxies; however, the intervention had a clinically meaningful effect on the CSO's coping as measured by their emotional consequences, anxiety, depression, and relationship satisfaction. Several methodological issues are discussed in relation to this RCT. For transparency and for better pooling of data, we also published the raw data, including all measured outcomes together with the R scripts used to analyze the trial. The data and scripts can be downloaded from https://osf.io/awtg7.

Psychotherapy researchers face significant challenges, and there is a great need for high-quality psychotherapy trials, a better appreciation of the methodological issues, and more transparent reporting practices. Hopefully, improvements to psychotherapy research will follow, and that these improvements will improve clinical practice and reduce the mental health burden in general.

List of Scientific Papers

I. **Magnusson, K.**, Andersson, G., & Carlbring, P. (2018). The consequences of ignoring therapist effects in trials with longitudinal data: A simulation study. *Journal of Consulting and Clinical Psychology*, 86(9), 711–725. https://doi.org/10.1037/ccp0000333. **Code and data:** https://osf.io/egj4m/

II. **Magnusson, K.**, Nilsson, A., & Carlbring, P. (2019). Modeling Longitudinal Gambling Data: Challenges and Opportunities *[Preprint]*. https://doi.org/10.31234/osf.io/uvxk2. **Code**: https://osf.io/6pbgv/

III. **Magnusson, K.**, Nilsson, A., Andersson, G., Hellner, C., & Carlbring, P. (2019). Level of Agreement Between Problem Gamblers' and Collaterals' Reports: A Bayesian Random-Effects Two-Part Model. *Journal of Gambling Studies*. https://doi.org/10.1007/s10899-019-09847-y. **Code:** https://osf.io/ec5b9/

IV. **Magnusson, K.**, Nilsson, A., Andersson, G., Hellner, C., & Carlbring, P. (2019). Internet-delivered cognitive-behavioral therapy for significant others of treatment-refusing problem gamblers: A randomized wait-list controlled trial. *Journal of Consulting and Clinical Psychology*, 87(9), 802–814. https://doi.org/10.1037/ccp0000425. **Code and data:** https://osf.io/awtg7/

Other Relevant Publications

Magnusson, K., Nilsson, A., Hellner Gumpert, C., Andersson, G., & Carlbring, P. (2015). Internet-delivered cognitive-behavioural therapy for concerned significant others of people with problem gambling: study protocol for a randomised wait-list controlled trial. *BMJ Open*, 5(12), e008724. https://doi.org/10.1136/bmjopen-2015-008724

Nilsson, A., **Magnusson, K.**, Carlbring, P., Andersson, G., & Hellner Gumpert, C. (2016). Effects of added involvement from concerned significant others in internet-delivered CBT treatments for problem gambling: Study protocol for a randomised controlled trial. *BMJ Open*, 6(9), e011974. https://doi.org/10.1136/bmjopen-2016-011974

Nilsson, A., **Magnusson, K.**, Carlbring, P., Andersson, G., & Hellner Gumpert, C. (2018). The development of an internet-based treatment for problem gamblers and concerned significant others: A pilot randomized controlled trial. *Journal of Gambling Studies*, 34(2), 539–559. https://doi.org/10.1007/s10899-017-9704-4

Contributions to Open Source Software

Magnusson, K. (2018). powerlmm: Power analysis for longitudinal multilevel models. R package version 0.4.0 https://CRAN.R-project.org/package=powerlmm

Contents

Abbreviations

Abbreviation	Term
ATE	Average treatment effect
CBT	Cognitive-behavioral therapy
CRAFT	Community reinforcement and family training
CSO	Concerned significant other
CI	Confidence interval
DGP	Data-generating process
GEE	Generalized estimating equation
GLMM	Generalized linear mixed-effects model
ITT	Intention-to-treat
ICC	Intraclass correlation
JM	Joint-model
LRT	Likelihood ratio test
LMM	Linear mixed-effects model
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
OSF	Open science framework
PM	Pattern-mixture
PDT	Psychodynamic therapy
RCT	Randomized controlled trial
RMSE	Root-mean-square error
SE	Standard error
SEK	Swedish krona
TLFB	Timeline follow-back

Chapter 1

Introduction

> *"Like all therapists, I personally experience an utter inability not to believe I effect results in individual cases; but as a psychologist I know it is foolish to take this conviction at face value. In order to bring about the needed research, it will probably be necessary for therapists and administrators to get really clear on this point: Our daily therapeutic experiences, which (on good days!) make it hard for us to take Eysenck seriously, can be explained within a crude statistical model of the patient-therapist population that assigns very little specific 'power' to therapeutic intervention. If the majority of neurotics are in 'unstable equilibrium' and hence tend to improve under moderately favorable regimes, those who are in therapy while improving will be talking about their current actions and feelings in the sessions. Client and therapist will naturally attribute changes to the therapy."*
>
> – Paul E. Meehl (1955)

Over the last couple of decades evidence-based psychotherapies have flourished, and there are now therapies that are well-established for a wide range of problems (Nathan & Gorman, 2015). At the same time the mental-health burden is still enormous, and challenges to the dissemination of treatments are substantial (Holmes et al., 2018). While some celebrate the enormous achievements made by the evidence-based therapy movement, others are more concerned about the quality of the evidence.

In the clinical sciences, especially in the biomedical ones, the issues of research quality has had some vocal critics. In 1994 Doug Altman published his famous editorial *The scandal of poor medical research* which begins with the widely spread phrase "We need less research, better research, and research done for the right reasons" (Altman, 1994). In an attempt to improve the reporting of clinical trials, the CONSORT statement was published 1996. The threat of various biases received increased attention ten years

later when Ioannidis (2005) published the highly influential and discussed paper *Why most published research findings are false*. Similarly, Chalmers & Glasziou (2009) estimated that 85% of the investments in biomedical research are wasted, due to a combination of researchers focusing on outcomes that are not important to patients and clinicians, flawed research designs, studies never being published, and an underreporting or lack of transparency when findings are reported. The Lancet published a special series on increasing value and reducing research waste, where Macleod et al. (2014) argues that initially promising findings that fail to improve healthcare outcomes are the norm.

In psychology, the most active discussion regarding research quality has been by experimental and social psychologists. Perhaps, most famously through the reproducibility project (Open Science Collaboration, 2015), where a large team of researchers tried to replicate 100 psychological experiments. In a large part of the replication attempts they found much weaker evidence compared to the original investigation. Likewise, in several influential papers, different authors have pointed out that several questionable research practices could contribute to untrustworthy studies (Gelman, 2013; Simmons, Nelson, & Simonsohn, 2011), and highlighted the wrong incentives at play (Bakker, van Dijk, & Wicherts, 2012; Nosek, Spies, & Motyl, 2012). Many of these concerns are likely behind the growing open science movement, consisting of people advocating for a more transparent and open science (Nosek et al., 2012; Wallach, Boyack, & Ioannidis, 2018).

How is the situation in clinical psychology? Are our results more robust and our trials more transparently reported? The replicability crisis has been frequently discussed in psychology in general, and as mentioned, especially in social psychology. However, clinical psychology has, according to some, been uninterested in participating in the discussion (Hengartner, 2018; Tackett et al., 2017). Although the focus on open science might be new, in the subfield of psychotherapy research, an active debate has been going on for decades about the quality of psychotherapy research. In 1952 Eysenck published a review claiming that psychotherapy was ineffective and that change could largely be attributed to spontaneous remission (Eysenck, 1952), which spurred a heated discussion.

Back when Eysenck published his critique the evidence for (and against) psychotherapy was mostly based on anecdotal clinical observations. Now several decades later, both clinicians and researchers act as if psychotherapy has clearly been established as efficacious in gold-standard and high-quality randomized controlled trials (RCTs). Although substantial gains in knowledge have been made, many issues remain unsolved, and there are many reasons to be skeptical of the current quality of the evidence. In the first part of this thesis, I give an overview of the broader discussion about the contemporary issues that concern the scientific investigation of psychotherapies, and

threaten the validity of psychological treatment research. After the broader overview, I present a more detailed background to the two major issues investigated in Study I and II. Chapter 3 covers therapist effects and Section 4.2 semicontinuous gambling data. In Chapter 4, I cover gambling disorder and especially research that focuses on the concerned significant others (CSOs) of problem gamblers.

Skärholmen, Oktober 2019

Chapter 2

General Issues in Psychotherapy Research

> *"So why describe them again? Our response was that many in the field of psychotherapy research are not aware of these fundamental problems, and are very much internally oriented with little knowledge about major developments in the methodologies in the broader biomedical field."*
>
> – Cuijpers, Karyotaki, Reijnders, & Ebert (2018, p. 1)

Psychotherapy researchers have generally focused on three questions: what treatment works, for whom does it work, and how does it work? These questions go back to the strategic outcomes proposed by Paul (1967): "*What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" (p. 112). In the remainder of this chapter, I will focus on these three questions and explore specific issues that are threats to the conclusions we draw from the literature.

2.1 How Well Do Treatments Work? Issues Estimating Treatment Effects

Claims that psychotherapies are efficacious for a wide range of problems are ubiquitous (Cuijpers et al., 2016; Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012). Some even argue that the evidence of efficacy and specificity are so strong that some psychotherapies are better called *psychological treatments* (Barlow, 2004). In this section, I will cover different biases and flaws that are threats to the quality of the evidence. As noted by Leichsenring et al. (2017), although, biases are recognized in the literature, several important biases are not controlled for, and as stated by Cuijpers et al. (2018a), psychotherapy researchers tend not to be aware of the fundamental problems.

2.1.1 Defining Treatment Effects

The psychotherapy literature is filled with reports that interchangeably use the terms: treatment outcome, treatment response, treatment effect, and outcome of treatment. Some authors use these terms when referring to the observed outcomes after treatment, and others when referring to the comparative difference between the outcomes in a control condition versus a treatment condition. Clearly, the outcome after treatment and the relative difference between the two groups are two different targets. Thus, before discussing specific biases, it is worthwhile to define what we mean by "treatment effects" and what trials try to estimate. In the causal inference literature, the treatment effect is often defined using potential outcomes (Rubin, 1974). For a single patient that treatment effect is,

$$\Delta_i = Y_i(1) - Y_i(0).$$

Here $Y_i(1)$ represents the potential outcome if the patient enters treatment, and $Y_i(0)$ the potential outcome if the patient decides to continue without treatment—representing the "what-if" scenario of what would have happened had the patient not decided to enter therapy. Thus, the causal effect is the relative efficacy of a treatment outcome compared to a counterfactual outcome. It should be evident that it is fundamentally impossible to observe this effect—we cannot simultaneously observe a patient's outcome after treatment and the treatment-free outcome. This fact is called *the fundamental problem of causal inference* (Holland, 1986; Rubin, 1974).

Instead, the usual causal estimand, i.e. the target of efficacy studies, is the average treatment effect (ATE), which under random assigment can be shown to be equal to (Holland, 1986),

$$\begin{aligned}
\text{ATE} = \Delta &= \mathrm{E}[Y_i(1) - Y_i(0)] \\
&= \mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)] \\
&= \mathrm{E}[Y_i(1) \mid Z_i = 1] - \mathrm{E}[Y_i(0) \mid Z_i = 0] \\
&= \mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0],
\end{aligned}$$

where Z indicates the outcome of the random treatment allocation with $Z = 0$ represents being assigned to the control condition and $Z = 1$ being assigned to the treatment condition. This equation might look more complicated than it is. It simply tells us that we can estimate the *average* treatment effect in the population by comparing the average outcome in the treatment group to the average outcome in the control group. Thus, the often-criticized focus on group-level effects in RCTs is more caused by a necessity rather than an uninterest in the individual causal effects. Moreover, it does not follow that we must believe that the treatment effect is the same for all patients—it is still possible

that individual-level treatment effects vary. However, we cannot directly observe those effects, so the ATE is our best guess what effect a treatment would have for a patient picked at random from the population. It is important to remember this "fundamental reality of causal analysis" (p. 14, Morgan & Winship, 2014), because as we will see later, investigators frequently forget this and draw conclusions about individual-level effects that could never be observed.

However, the effect of randomization is not everlasting, and the effect is frequently broken in trials; participants can deviate from their assigned treatments, e.g., miss sessions, fail to complete homework assignments, or simply drop out from the treatment. For those reasons, most RCTs use the *intention-to-treat* (ITT) comparison, where participants are analyzed according to how they were randomized. The ITT comparison is different from the ATE in that we are now evaluating the impact of the random allocation, i.e., the impact of *offering* treatment (Hernán & Hernández-Díaz, 2012). The ITT and ATE effects would be equivalent under perfect adherence and no dropout. Thus, from the point of view of trying to evaluate the efficacy of a treatment, psychotherapy trials could aptly be described as "broken trials" (Barnard, Frangakis, Hill, & Rubin, 2003), or as "longitudinal studies with baseline randomization" (Toh & Hernán, 2008).

Some investigators recognize the limitations of the ITT comparison in relation to estimating the efficacy of a treatment. Unfortunately, they often perform a "naive" per-protocol analysis to estimate the effect among those who adhered to the treatment. However, adherence is not experimentally manipulated; hence, the sample in the per-protocol analysis will likely be partially self-selected and prone to time-varying confounding, and valid inference would require dealing with those issues (Dunn, Maracy, & Tomenson, 2005; Hernán & Hernández-Díaz, 2012; Mohammad Maracy & Graham Dunn, 2011). Despite the challenges of estimating the causal effect among treatment completers, it is still worthwhile to remember that an ITT comparison might not be synonymous with efficacy in neither the mind of clinicians nor patients. Personally, either as a patient or a clinician, I would want to know the expected effect of an intervention if I actually complete it.

2.1.2 Comparators

There is an extensive disagreement regarding the choice of counterfactual in psychotherapy studies, i.e., the choice of the control condition. Trials investigating pharmacological therapies have the double-blind placebo RCT as the ideal—where the treatment effect is attributable to the active ingredients. For psychotherapy trials, the choice of comparison is not as straight-forward, and there is a large disagreement

among researchers. The disagreement can be summarised as: should non-specific effects be included in the causal treatment effect? Some would define the treatment effect as the effect that is above and beyond the effect of having contact with a warm and empathic therapist, and try to control for most of the non-specific effects (Gold et al., 2017; Guidi et al., 2018), using some type of "psychotherapy placebo." Proponents of the importance of common factors (non-specific) effects have called placebo controls in psychotherapy a flawed concept (Wampold, Frost, & Yulish, 2016), see Section 2.2.1 on page 11 for a background to the *common factors* theory. Much of the discussion is related to a difference in the view of what constitutes a treatment effect. From a common factors perspective it would be strange to try to control for the effect of the patient-therapist relation or expectation effects, and it would be conceptually hard to design an intervention where these are "placebo" components. Kirsch, Wampold, & Kelley (2016) writes: "in evaluating the efficacy of psychotherapy, the placebo effect cannot and should not be controlled" (p. 121), and Kirsch (2005) notes that "it is not clear how the effects of placebos are any less specific than any other psychologically produced effects." (p. 797).

However, proponents of specific effects claim that most of the therapy works by targeting and modifying specific processes related to the psychopathology. Naturally, from this point-of-view, investigators claiming that non-specific effects are the principal treatment mechanism should control for non-specific effects (Mohr et al., 2009).

Most investigators probably agree that an inert psychotherapy placebo condition that can be blinded is challenging to construct, which has lead to a discussion about what the appropriate comparator should be. It is possible that these challenges were instrumental in the adoption of waitlist controls—a comparator that has received ample criticism (Cristea, 2018). Several meta-analyses have provided evidence that a waitlist act as weak control condition (Huhn et al., 2014; Khan, Faucett, Lichtenberg, Kirsch, & Brown, 2012), with effects lower than expected by the natural course of spontaneous remission. This *nocebo* effect has been found in multiple meta-analyses (Cuijpers et al., 2016; Furukawa et al., 2014; Khan et al., 2012).

2.1.3 Biases, Was Eysenck Right?

There are multiple sources that could bias the estimates of treatment effects. Researchers' allegiance to the studied psychotherapy has frequently been brought up as a risk (Luborsky et al., 1999). Munder, Brütsch, Leonhart, Gerger, & Barth (2013) performed a meta-analysis of meta-analyses and found that the association between allegiance and outcome was, $r = .262$, $p = .002$. There are many plausible ways investigators' allegiance could bias treatment effects (cf., Luborsky et al., 1999; Coyne & Kok, 2014; Cuijpers

Figure 2.1: Allegiance and the outcome could share a common cause, reflecting the fact that both are influenced by the true efficacy of a treatment

& Cristea, 2016), they could: 1) pick a weaker control condition, 2) avoid publishing negative findings, 3) selectively report outcomes favoring their preferred treatment, 4) compare various different statistical models and pick the one that gives the most favorable results, 5) be more enthusiastic toward their prefered treatment and create stronger expectations in participants assigned to it, 6) recruit more competent therapists providing the preferred treatment, and 7) they can pick outcomes more responsive to their intervention. Naturally, these questionable practices need not be consciously performed by the investigator. Moreover, as Leykin & DeRubeis (2009) noted, the allegiance-outcome association might be a case of reverse causality, and that the causal effect might reflect "nature". Some researchers might have an allegiance to a treatment because it is actually superior. Similarly, allegiance and outcome could be concidered to share a common cause, as shown in Figure 2.1, thus, allegiance and outcome would not need to be causally realated. Moreover, it is possible that treatment effects from studies by investigators with high allegiance give more correct efficacy estimates, as they are more likely to be delivered by experts in the intervention (Gaffan, Tsaousis, & Kemp-Wheeler, 1995; Leykin & DeRubeis, 2009). Thus, bias would arise if there is differential competence between treatments which would need to be balanced to avoid bias (Hollon, 1999).

A real threat that can bias results are the unhealthy incentives in science, where you are not evaluated by the quality of your work but on the number of publications and ranking of the journals where you publish (Bakker et al., 2012). As Nosek et al. (2012) noted, there is "a disconnect between what is good for scientists and what is good for science." Indeed, this is also true for psychotherapy researchers and combined with allegiance it could lead to an unhealthy flexibility in the data analysis, and make investigators act as if a hypothesis was specified *a priori* when it was not, i.e., "hypothesizing after the results are known" (HARKing; Kerr, 1998). Even without allegiance, common problems in science are likely to apply to psychotherapy researchers; such as, the problems with outcome switching and selective reporting of studies that are not prospectively registered, enabling investigators to act as if a secondary outcome was the primary. Bradley, Rucklidge, & Mulder (2017) reviewed trials published between 2010 and 2014 in the five journals with the highest journal impact factor. Out of 112 trials, about 12% were prospectively and correctly registered. About half of the correctly and prospectively

registered trials showed signs of selective outcome reporting (7 out of 13 trials), leading to the authors to the stark conclusion:

> "We cannot currently have confidence in the results being reported in psychotherapy trials given there is no means of verifying for most trials that investigators have analyzed their data without bias" (p. 65).

Cybulski, Mayo-Wilson, & Grant (2016) reported similar numbers after reviewing 163 trials published in 2013: 15% were prospectively registered, and only two were prospectively registered and fully described their primary outcome. Azar, Riehm, McKay, & Thombs (2015) reviewed all trials published in one of the top clinical psychology journals, the *Journal of Consulting and Clinical Psychology*, and found a similarly small proportion of preregistrations (17%, 12/70 studies).

Concerns for biasing effects of these research practices should not be taken lightly. More recent meta-analyses shows that the effects of CBT are small to moderate in methodologically rigorous studies (Cuijpers et al., 2016, 2010b). There is even some evidence, that when focusing on high-quality trials the effect of psychotherapy is no longer clinically relevant (Cuijpers et al., 2014), and that publication bias most likely contributes to an overestimation of the treatment effect (Cuijpers et al., 2010a; Driessen, Hollon, Bockting, Cuijpers, & Turner, 2015), which in turn will have a negative impact on psychotherapies' replicability and trustworthiness (Leichsenring et al., 2017). For instance, Flint, Cuijpers, Horder, Koole, & Munafò (2015) meta-analysed 149 psychotherapy studies and came to the conclusion that the studies included much more significant findings than what would be expected. Recent reviews have also pointed out that non-financial conflicts of interest are a cause of concern both in the reporting of trials (Cristea & Ioannidis, 2018), and in systematic reviews (Lieb, Osten-Sacken, Stoffers-Winterling, Reiss, & Barth, 2016). The fallibility of meta-analysis is nothing new, Eysenck (1978) called it an "exercise in mega-silliness" and noted that it is hard to overcome the problem of "garbage-in-garbage-out" (p. 517).

2.1.4 What Should the Treatment Affect?

Even if we manage to estimate a treatment effect without bias, this might mean very little if the thing we are measuring is unimportant or lack validity. Over the years, there have been many debates about what outcomes we should focus on. Some have questioned the *symptom-reduction model*, and argued that the focus should be on transsyndromal symptom reduction such as resilience and social participation (Os, Guloksuz, Vijn, Hafkenscheid, & Delespaul, 2019).

Selecting a relevant outcome domain is only the first issue—deciding how to measure

the target is even more challenging. In psychotherapy trials, it is common that an outcome is measured using a patient- or clinician-rated scale where some aggregate score, such as a sum score, is used as the outcome (Ogles, 2013). This would perhaps be reasonable if all items measured the same thing, and the items' importance was appropriately weighted. However, there is also a growing discussion regarding if we should instead focus on individual symptoms and their interactions. Some argue that mental disorders are well described by the frequent view that a latent brain disorder is a common cause of the observed symptoms and that a valid scale is a unidimensional measure of this construct. Others have proposed that disorders are best viewed as to have formed from causally interrelating symptoms (Borsboom & Cramer, 2013). From this view, the usual method of assessing treatment outcome by analyzing sum scores from a questionnaire risks obfuscating important insights. For example, we could hold the view that a person with a gambling problem gambles too much because there is a latent neurological dysfunction that could be explained by neuroscientists—and that such a dysfunction could be targeted psychopharmacologicaly. Alternatively, we could focus on the actual symptoms and try to identify important symptoms and potential causal pathways; a view that most behavioral therapists would recognize as important. For instance, a simplified chain of symptoms could be: stress $\rightarrow$ gambling $\rightarrow$ debt $\rightarrow$ more stress $\rightarrow$ gambles to solve debt $\rightarrow$ even more stress $\rightarrow$ and so on. For this patient, a good treatment outcome might be if we could help them reduce their stress and gambling losses. For another patient, the losses might not be the central problem, but that they spend too much time occupied with gambling so that other areas of their life are negatively impacted.

Things get even more difficult when trying to measure, for instance, depression, where there is not a clear problematic behavior in the same way as for problem gambling (Bagby, Ryder, Schuller, & Marshall, 2004; Fried, 2017). There is a clear challenge in verifying that a scale has the same meaning both in different groups and over time (Reise & Waller, 2009), if not, an outcome of 10 would mean different things for different people or at posttest compared to baseline. Furthermore, if the scale is multidimensional so that items measure different constructs, then the treatment might affect one of these constructs, which means that important improvement could be lost in the aggregated score (Bagby et al., 2004).

Much more could be said about the measurement issues in psychotherapy studies. However, true to the spirit of most published research, I will recognize measurement issues as hugely important but focus very little on them in this thesis.

2.2 How Do Treatments Work? Issues with Research on Mechanisms and Mediators

If we acknowledge that standard psychotherapy trials struggled to provide robust evidence both for a treatment's efficacy and its specificity, this leads to the questions *how* do treatments work then? The discussion about how psychotherapy leads to change has been debated for a long time. Before covering the issues with studying mechanisms, I will first provide some background on the contrasting views of proponents of common factors and specific effects.

2.2.1 Common Factors

Rosenzweig (1936) proposed that the important mechanisms are factors common to all psychotherapies. Such common factors are often assumed to be the therapeutic alliance, expectations, and empathy (Wampold & Imel, 2015). The common factors theory makes the strong claim that benefits from psychotherapies operate via mechanisms that are common to all psychotherapies.

The modern common factors theory that has been the most influential is the *contextual model* (Wampold & Imel, 2015). The contextual model proposed three pathways by which psychotherapy works: 1) the real relationship, 2) creating expectations, and 3) health-promoting actions (Wampold, 2015). Thus instead of targeting specific ingredients, the contextual model sees psychotherapy as a social healing practice. Wampold (2015) writes:

> "the contextual model provides an alternative explanation for the benefits of psychotherapy to ones that emphasize specific ingredients that are purportedly beneficial for particular disorders due to remediation of an identifiable deficit" (p. 270).

When therapy starts, the patient and the psychotherapist form a bond, and it is crucial to establish an engaging and trusting relationship. The "real relationship" that forms is thought to be healing in and of itself; the patient forms a human connection with an emphatic individual that cares for their well-being. The psychotherapy then creates expectations by providing a plausible explanation for the patient's suffering and offering a rationale for how to overcome the difficulties. These expectations increase the patient's feelings of self-efficacy that they can solve their problems. Thus, in the contextual model, it is critical that the patient and the therapist believes in the explanation and the proposed solutions. The third path, health-promoting actions, include specific ingredients. However, in the contextual model, the therapeutic

ingredients are thought to work because they influence the patient to do something that is health-promoting. Perhaps the most striking difference to the view of proponents of specific ingredients is shown in this passage by Wampold & Imel (2015):

> "What is important for creating expectations is not the scientific validity of the theory but the acceptance of the explanation for the disorder, as well as therapeutic actions that are consistent with the explanation … The causes of mental disorders are notoriously difficult to determine … and for the sake of creating expectation are irrelevant. If the client believes the explanation and that engaging in therapeutic actions will improve the quality of their life or help them overcome or cope with their problems, expectations will be created and will produce benefits." (p. 59)

In an often referred to pie chart, Lambert (1992) state that 30% of the change can be attributed to common factors and 15% to specific effects. Although, Wampold & Imel (2015) notes that "[s]uch attempts are flawed for several reasons. First, partitioning variability in outcomes to various sources assumes that the sources are independent, which they are not" (p. 256), yet they still include a similar table that is also often referred to. For instance, Wampold & Imel (2015) claim that alliance explains 7.5% of the variance, whereas the difference between psychotherapies explains less than 1% of the variance. However, in my opinion, these types of presentations are highly misleading and are of very little value. As covered previously, the target if a trial is not to explain variance in outcomes, it is to estimate a causal treatment effect. It is important to remember that the treatment effect is comparative, and correlating alliance with, e.g., the outcome after treatment does not mean that 7.5% of the treatment *effect* is explained by the alliance. Thus, even though it is possible to convert both effect sizes to a percentage, one still has to remember that two fundamentally different quantities are being compared—and that such a comparison makes no sense.

From the common factors perspective, the specific ingredients are not what drives the treatment effect. Closely related to this proposition is the "dodo bird verdict", which says that all *bona fide* psychotherapies *intended* to be therapeutic have similar effects. The name "Dodo bird" originates from a quote from *Alice in Wonderland* used by Rosenzweig (1936): "At last the Dodo said, 'Everybody has won, and all must have prizes' " (p. 412). Proponents of common factors argue that instead of focusing on specific effects, researchers should try to understand how to maximize the effects of therapeutic alliance and expectation effects (Kirsch et al., 2016; Wampold & Imel, 2015).

2.2.2 Specific Effects

In some therapy traditions, specific effects are thought to be the main factor by which the treatment works (e.g., in CBT). From this perspective, a specific ingredient targets a specific dysfunction that is thought to have an important effect on the psychopathology. For instance, in a behavior therapy for a person with an addictive disorder, such as gambling, the therapist might theoretize that the patient is engaging in problematic gambling because they are depressed and that gambling offers a way of escaping these feelings. This process is then specifically targeted in therapy by encouraging the patient to engage in other activities that also reduce the depressive feelings, but without the negative consequences caused by gambling. In other cases, it can be conceptualized that the gambling is caused by a lack of control of impulses, and that the treatment should target that dysfunction.

Proponents of specific effects do not generally ignore the importance of common, or nonspecific, factors. However, it is thought that the specific ingredients have important effects and that they work by targeting the psychopathology (Barlow, 2004). Barlow (2004) notes that these treatments "are specifically tailored to the pathological process that is causing the impairment and distress" (p. 873). True to the medical model, these techniques should preferably be derived from basic sciences, such as behavioral and cognitive sciences.

2.2.3 Studying Mechanisms

One might wonder why, after decades of psychotherapy research, is there no irrefutable evidence for either specific *or* non-specific effects? In this section, I will cover studies of mechanisms and the monumental, and often overlooked, challenges associated with such investigations.

Much hope has been put in improving psychotherapies by better understanding its mechanisms (Holmes et al., 2018; Kazdin, 2011; Nock, 2007). Barlow, Bullis, Comer, & Ametaj (2013) writes: "Progress in identifying and confirming mechanisms of actions will be one of the most efficient methods for improving treatment efficacy" (p. 14). Naturally, it is a worthwhile goal to try to identify a specific mechanism that drives the treatment effect. If such a causal process could be identified, the hope is that psychotherapies could be improved by distilling the efficacious mechanism of change (Barlow et al., 2013). Alternatively, if we modify a treatment, for instance, from a face-to-face setting into a smart-phone app or an internet treatment—understanding critical mechanisms could help, especially if the therapeutic alliance is shown to be a vital component (Kazdin, 2011). In *The Lancet Psychiatry Commission on psychological treatments*

research in tomorrow's science Holmes et al. (2018) writes: "Research on these mechanisms has considerable scope to facilitate treatment innovation" (p. 237), and they call for greater collaboration between basic researchers and clinical researchers—reminiscing the seminal mechanistic studies in the 1950s and 1960s that are now cornerstones of today's evidence-based therapies.

From a slightly different perspective, Hofmann and Hayes have coined the term *process-based therapy* (Hayes et al., 2018; Hofmann & Hayes, 2018). As an alternative to targeting distinct syndromes via specific treatments, they argue for an expansion of the strategic outcomes proposed by Paul (1967). According to them, psychotherapy researchers should focus on core evidence-based processes and move away from treatment packages. In their view, the investigation of mediators and moderators will be crucial in this process—which should lead to a rise in mediation and moderation studies.

However, even if we could rule out non-specific effects via placebo control conditions—or use a dismantling study to show that removing a component impacts the outcome—could we then deduce that the psychotherapy works through specific ingredients? That answer is that it depends on what assumption you are willing to make. Such designs could point towards essential components, but they would not necessarily explain *how* the mechanism works (Kazdin, 2007). As I will cover in this section, the strong assumptions that are needed to identify a mechanism either via statistical analysis or via experimental manipulation are largely underappreciated by psychotherapy researchers.

2.2.4 Mediation Analysis

Statistical mediation analysis is likely the most common method used when investigating mechanisms. A mediator is an intervening variable that transmits some or all of the effect of the treatment on the outcome (Baron & Kenny, 1986). The mediated effect, often called the indirect effect, gives us the expected change in the outcome caused by the treatment's effect on the mediator, while the direct effect of treatment is the total effect of all the remaining causal mechanisms. From a CBT-perspective, an apparent mediator would be adherence to homework assignments, where one would expect that part of the treatment effect would be transmitted via the completion of homework assignments. The relationship between homework and treatment effects has received a lot of attention (Burns & Spangler, 2000; Driessen & Hollon, 2010). Common factor proponents, on the other hand, have focused most attention on the allegiance between patient and therapist (Wampold, 2015), as stated by Wampold (2005), "[t]he working alliance is the ubiquitous common factor that has been claimed to be causal to outcomes" (p. 195).

The most influential paper on mediation analysis is probably by Baron & Kenny (1986),

which at the time of writing has over 80,000 citations on *Google Scholar*. Figure 2.2 shows a causal diagram depicting three scenarios involving either homework adherence or alliance. The main challenge is that the mediator is not randomized, and therefore it is likely that there is one or several variables that influence both the mediator and outcome. In (a), the standard mediation figure (Baron & Kenny, 1986) is shown, except that there is a potentially unknown variable influencing both the mediator and the outcome. When evaluating a causal diagram, the important question to ask is what arrows are missing, and can such an assumption be justified. In the case of homework completion, one could easily argue that the group of patients that complete more homework would also have had relatively better outcomes had they been assigned to the control group, i.e., their prognosis is better which is caused by, for instance, age and education. If we have measured both age and education, then we can adjust for them, and the effect of homework can be identified. In (b), the causal model is instead that it is allegiance that mediates the outcome, and homework completion is just an effect of better allegiance and not causally related to the outcome. Thus if we used homework as a mediator, we would wrongly conclude that there is an effect of homework completion, while in fact it is confounded by allegiance. Of course, the relationship between allegiance and outcome could also be confounded. In (c), homework does mediate the treatment effect, but the effect is influenced by a known and measured confounder (baseline functioning). However, there is also an unknown confounder that influences both alliance and outcome. In this scenario, we would identify the true effect of homework if we adjust for baseline functioning and do *not* include allegiance in the model, as allegiance would be a collider and bias the results. These are just three hypothetical examples of how easily it is to draw the wrong conclusions from these types of observational data that result from an RCT.

In the notation of the potential outcomes framework the indirect effect for a single patient (Emsley, Dunn, & White, 2010; Imai, Tingley, & Yamamoto, 2013), is written as,

$$\text{indirect effect} = Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

and the direct effect of the treatment is,

$$\text{direct effect} = Y_i(1, M_i(t)) - Y_i(0, M_i(t)).$$

$M_i(1)$ is the level of the mediator under the treatment and $M_i(0)$ under the control. Thus, the causal estimand representing the mediated effect for a patient can be written as: their outcome under the treatment at the level of the mediator that would occur if they were assigned to the treatment *compared* to the outcome if the mediator was set to the level that would occur had they received the control intervention. Alternatively, as Imai et al.

(2013) described it: "what change would occur to the outcome if we change the mediator from the value that would realize under the control condition … to the value that would be observed under the treatment condition … while holding the treatment status at *t*" (p. 8).

Considering how complex psychotherapy processes are one would assume that researchers have paid careful attention to the causal assumptions underpinning mediation analysis. Unfortunately, that is not the case, and the critique from methodological experts has been vicious. One of the most basic sources of confusion, seems to be how to even define treatment effects, and you can often see RCTs where authors ignore the control group and look at relationships only in the treated group. As noted by Dunn & Bentall (2007) "[t]his approach is based on the mistaken assumption that the outcome of treatment (i.e., the outcome following treatment) is a measure of that treatment's effect." (p. 4742) and they continue:

> "Although there is an enormous methodological literature on the estimation of the effects of mediators (particularly, in psychology … most of it completely ignores the technical challenges raised by measurement errors in the proposed mediators and by potential hidden confounding of mediator(s) and outcome." (p. 4743).

The fact that (psychotherapy) researchers consistently make this unrealistic, and often unstated and unjustified, assumption is also noted by Kenny (2008) who write: "all too often persons conducting mediational analysis either do not realize that they are conducting causal analyses or they fail to justify the assumptions that they have made." (p. 356). Similarly, Bullock, Green, & Ha (2010) writes, "[t]his warning has been issued before by those who write about mediation analysis … but it seems to have escaped the attention of the mainstream of the discipline" (p. 551). Dunn & Bentall (2007) put it a bit more harshly:

> "The assumptions concerning the lack of hidden confounding and measurement errors are very rarely stated, let alone their validity discussed. One suspects that the majority of investigators are oblivious of these two requirements. One is left with the unsettling thought that the *thousands of investigations of mediational mechanisms in the psychological and other literatures are of unknown and questionable value.*" (p. 4743, italics added).

Their critique is unpleasant but, unfortunately, true in my experience. Considering that Judd & Kenny (1981), noted the shortcomings with the classical mediational model in 1981, perhaps, some negativity from statisticians is understandable. Within the context of psychotherapy, most texts on mediation analysis focus on the traditional approach.

However, there are some papers that includes examples of how instrumental variables and other statistical techniques can be used to help identify causal effect in the presence of both measurement error and confounding (Dunn & Bentall, 2007; Mohammad Maracy & Graham Dunn, 2011; Preacher, 2015; Valente, Pelham, Smyth, & MacKinnon, 2017).[1]

2.2.5 Experimental Manipulation of the Mediator

Psychotherapy researchers often mention that experimental manipulation of the mediator would provide strong evidence for a causal mechanism (Kazdin, 2007). For instance, Holmes et al. (2018) reasons that "[s]howing that experimental manipulation of a proposed mechanism leads to symptom change is a powerful method for validation" (p. 244). Technically, experimental manipulation of the mediator provides evidence for the $M \rightarrow Y$ path and not the $Z \rightarrow M \rightarrow Y$ path. Even if we can also show that the treatment has an impact on the mediator, $Z \rightarrow M$, this is not sufficient to claim that the treatment effect is transmitted via the mediator. Even under this scenario, where both the treatment assignment and the mediator are experimentally manipulated, strong assumptions are needed in order to interpret this as evidence for a causal indirect effect (Imai et al., 2013)—and as we shall see it is not easy to claim that the required assumptions hold for psychotherapies.

Both Imai et al. (2013) and Bullock et al. (2010) clearly details the challenges of studying mediators using experimental manipulation. Especially relevant to psychotherapy research is the issue that investigators need to show that the manipulation only affects the proposed mediator and not other variables. Moreover, researchers also need to show that the experimental manipulation does not influence the mediator differently and in different individuals compared to how the treatment is assumed to shift the mediator. This is important since the aim is to study how the *treatment* naturally influences the outcome via the mediator, if the experimental manipulation achieves this by different means and in different individuals then it would be hard to generalize this causal relationship to the original treatment setting. In a psychotherapy trial, it becomes conceptually challenging to imagine how one would manipulate the mediator while holding the treatment constant. Even if we were to successfully experimentally manipulate, say, homework adherence or therapist alliance, it would be incredibly difficult to claim that the treatment is not also changed. We would need to assume that this manipulation of the mediator has no direct effect on the outcome. This means that a patient's treatment outcome would be assumed to be the same no matter if the mediator (M) takes on value $M = m$ naturally or by experimental manipulation. Which means

[1] I have posted R code and some empirical examples of how confounding and measurement error leads to bias https://rpsychologist.com/adherence-analysis-IV-brms and https://rpsychologist.com/mediation-confounding-ME

that if we manipulate homework adherence patients must behave similarly under this manipulation as if they had chosen the level of adherence naturally. Are we willing to believe that a patient that naturally form a weak allegiance to the therapist behave the same and have identical outcomes as a patient, that is somehow, "experimentally manipulated" to have weak allegiance to the therapist? Imai et al. (2013) are explicit about the importance of this assumption (the "consistency" assumption):

> "The importance of assumption 3 cannot be overstated. Without it, the second experiment provides no information about causal mechanisms (although the average causal effect of manipulating the mediator under each treatment status is identified). If this assumption cannot be maintained, then it is difficult to learn about causal mechanisms by manipulating the mediator" (p. 12)

From a clinical point of view, causal heterogeneity is easy to imagine. Most bona fide psychotherapies consist of many different tools and hypothesize multiple different pathways. As a clinician, you try to conceptualize each participants problem and propose solutions tailored to this individual—with some you focus more on the relationship and what happens in the therapy room, with others exposure-based exercises in the real world seems more fitting. Obviously, we cannot directly observe such causal heterogeneity.

2.2.6 The Dodo Bird and the Absence of Evidence Fallacy

As noted previously, identifying mechanisms is incredibly hard, and most published papers are strictly correlational. Thus, empirical evidence for both common factors and specific effects are very limited. However, common factor proponents mostly point to the fact that differences between therapies tend to be small (Mulder, Murray, & Rucklidge, 2017; Wampold & Imel, 2015), and that this Dodo bird conjecture is evidence of the importance of common factors, or as stated by Wampold & Imel (2015):

> "Evidence consistent with Rosenzweig's claim of uniform efficacy—commonly referred to as the Dodo bird effect—is typically considered empirical support for the conjecture that common factors are the efficacious aspect of psychotherapy" (p. 114)

There are several problems with the Dodo Bird verdict—in addition to the fact that it says very little about mechanisms even if the proposition was true. A problem with the common factors and dodo bird argument is that unbiased effects of efficacy are tough to observe. As noted earlier, ITT effects can be quite poor estimates of the causal treatment effect. Indeed, it would be hard to argue that two psychotherapies produce similar effects

if most patients simply never shown up to treatment. Clearly, as a patient, I would want to know: what is the expected benefit if I show up to the treatment and complete it? However, the next section will show that even under perfect adherence and no missing data, it would be wrong to claim that a non-significant statistical test demonstrates that two treatments have equivalent effects.

2.2.7 Non-inferiority and Equivalence Studies

Many treatment modalities are inferred to be equally efficacious based on a statistical misunderstanding. This fallacy was famously summarised by Altman & Bland (1995) as: "Absence of evidence is not evidence of absence" (p. 485). Many psychotherapies have been deemed equally effective based on failing to reject the null hypothesis, i.e., a test of $H_0 : d = 0$, versus the alternative $H_a : d \neq 0$. A non-significant result (usually $p > 0.05$) does not warrant the conclusion that the null hypothesis of no difference is supported. If traditional significance tests could be used in this way, it would mean that a study with a *smaller* sample size would more often find evidence for the equivalence of two treatments, compared to a study with a larger sample size. This is well known in the methodological literature, where the appropriate test would be either a non-inferiority or equivalence test (Greene, Morland, Durkalski, & Frueh, 2008; Piaggio et al., 2006; Wellek, 2010). Non-inferiority and equivalence test have increasingly been used by psychotherapy researchers (e.g., Leichsenring et al., 2018; Steinert, Munder, Rabung, Hoyer, & Leichsenring, 2017), e.g., when comparing PDT vs. CBT (Driessen et al., 2013), or when comparing an internet-delivered treatment versus a face-to-face treatment (Lappalainen et al., 2014).

Simplified, we can say that the classical test is performed by checking if a confidence interval (CI) includes zero or not, whereas, a non-inferiority test implies testing if the lower end a CI is above $-\Delta$, i.e., it tests if the new treatment is at least not worse than the old gold-standard treatment by $-\Delta$ ($H_0 : d \leq -\Delta$). An equivalence test checks if the CI of the treatment effect falls within $[-\Delta, \Delta]$, i.e. we test if the difference is not larger than $\pm\Delta$ ($H_0 : d \leq -\Delta$ or $d \geq \Delta$).[2] Remembering that an equivalence test is essentially trying to squeeze a CI within a small region, we can also see that equivalence tests generally require large sample sizes. For example, an investigator planning to test if two interventions are equivalent using a t-test with $\Delta = 0.2$, would need approximately 500 participants *per* group and that is assuming there is no true difference—if a small but clinically meaningless effect exists the sample size would need to be even larger (c.f., Julious, 2004).

[2] I have created an interactive explanation showing the difference between superiority, non-inferiority, and equivalence test https://rpsychologist.com/d3/equivalence/

Non-inferiority and equivalence trials are not without their challenges. The choice of Δ is critical and an ongoing topic of discussion among psychotherapy researchers. In a recent meta-analysis, Steinert et al. (2017) concluded that (standardized) margins range from 0.24 to 0.6 and therefore adopted a margin $\Delta = 0.25$. However, Rief & Hofmann (2018) argued that such a "… threshold is inflationary, hiding clinically meaningful differences that might exist." (p. 1393), and went so far as to recommend that Δ should be "90% of the expected effects of the first-line treatments (e.g., a threshold SMD of ± 0.05, if the uncontrolled effect size is expected as SMD = 0.50)." Clearly, a $\Delta = 0.05$ will protect against degradation; however, as noted by Leichsenring et al. (2018) this would require 6,281 particpats per arm to reach 80% power.

The choice of Δ is not the only issues with non-inferiority and equivalence studies. As mentioned earlier, ITT analyses tend to underestimate the true efficacy of a treatment and are thus viewed as conservative. For the very same reason, ITT analyses are nonconservative with regards to testing non-inferiority or equivalence (Hernán & Hernández-Díaz, 2012). Thus, one can argue that these designs are less robust to being influenced by different biases; such as allegiance, non-adherence, or missing data.

2.3 For Whom Does the Treatment Work?

An alternative view to focusing on mechanisms and basic science in order to improve patient outcomes is to instead better match patients to the treatments. From this perspective, the way forward is either better treatment selection using predictive models or more personalized treatments. In this section, I will cover some of the challenges and misunderstandings that apply to these questions.

2.3.1 Identifying Treatment Responders

I would assume that most clinicians believe—as I do—that different patients benefit differently from the treatment, i.e., that there exist treatment effect heterogeneity. With some patients, you feel that there is very little progress, and with others that the improvements are substantial. There have been many papers arguing that a noticeable proportion of patients do not respond to treatment. For instance, Cuijpers, Karyotaki, et al. (2014) writes that 48% of depressed patients receiving psychotherapy respond to treatment (with "response" defined as a 50% reduction of symptoms). Researchers constantly invent new ways to solve the "problem" of treatment non-responders, such as: 1) trying match patients to the right treatment (Cohen & DeRubeis, 2018), 2) employing an idiographic approach to tailor treatments to patients needs (Fisher, 2015; Hofmann, Curtiss, & McNally, 2016), 3) identifying biomarkers such as genetic markers

or brain imaging to predict non-responders (Insel, 2014). Clearly, trying to improve therapy outcomes is laudable; however, most of these efforts ignore the fact that we do not know how many or who responds to treatment. As I will cover in this section providing evidence for such variation in treatment responses is much harder than most think.

Figure 2.3 illustrates that the problem with identifying variation in treatment response is—again—that the individual-level effects are unobservable. By only looking at the patients' change over time, we get the impression that their response to the treatment varies substantially. However, as covered previously, we do not know how their outcomes would have looked had they not received treatment—it is possible that everyone benefited equally from the treatment. It should be evident that the baseline is not a valid counterfactual for inferring (individual) treatment effects. The standard RCT cannot tell us how many of the patients benefited from the treatment, or even estimate if there exists between-patient variance in treatment effects (Senn, 2004, 2016). Estimating the variance in patient's individual treatment effects would require a repeated cross-over design where patients are repeatedly randomized (Senn, 2016). It is highly unlikely that such a trial would be a feasible way of evaluating a psychotherapy.

Several journals require reporting some type of clinical significance. Often researchers report how many patients have a reliable improvement and include a normative comparison (Jacobson & Truax, 1991), or who improve by a certain percentage compared to their baseline measure. Kendall, Marrs-Garcia, Nath, & Sheldrick (1999) states that clinical significance answers the question: "[i]s the amount of change exhibited by an individual participant large enough to be considered meaningful" (p. 283). Clearly, these types of measures say very little about the clinical significance of a treatment effect. One would think that these metrics are intended to be descriptive and show that, e.g., a large proportion of patients still have residual symptoms. However, Jacobson, Roberts, Berns, & McGlinchey (1999) clearly had causal ambitions with their measure:

> "Jacobson and colleagues attempted to grapple with two limitations prevalent in statistical comparisons between groups of treated clients. First, such comparisons provide little or no information regarding the variability in *treatment response from person to person*. Group means, for example, do not in and of themselves indicate the proportion of participants who have improved or recovered *as a result of treatment*. Thus, statistical comparisons between groups shed little light on the proportion of participants in each condition who have benefited from the treatment" (p. 300, italics added)

Of course, we could calculate the proportion of patients in the treatment and control group that are in remission, or who have scores in the normal range. Then use the

relative comparison as the treatment effect; however, now we have just dichotomized a continuous outcome. It is not clear that such a dichotomization represents clinical significance any better than the average treatment effect using the continuous outcome. For instance, if we recruit a sample of patients with very high levels of symptom severity, a relatively large treatment effect can be found while zero of the patients are classified as recovered. Should we now conclude that the treatment had no clinically significant benefit?

A simple thought experiment can easily demonstrate the faulty logic behind using patients' change over time as the basis of inferring clinical significance. Take the patient in Figure 2.3 that deteriorated the most over time; this patient would be labeled as having a reliable deterioration and as a treatment non-responder. Now, imagine, as shown in the figure, that their—fundamentally unobservable—individual treatment effect is positive, meaning, that had they not received treatment their deterioration would have been even greater. Taken to the extreme, it is possible that the treatment kept them from committing suicide. Thus, even though the patient is still highly depressed, it would be strange to say that the treatment failed if the treatment actually *saved their life*.

2.3.2 Moderators and Personalized Psychotherapy

Variables that modify the treatment effect (moderators) are conceptually much easier to identify than mediators—but the search for moderators is not without its challenges. Research on moderators has a long history in psychotherapy research (Beutler, 1991; Kazdin & Blase, 2011; Paul, 1967).

One pitfall when identifying moderators is when researchers perform separate subgroup analysis and claim moderation when the treatment is significant in one group and not in the other (e.g., in men versus women). Although, my impression is that it is fairly well established among psychotherapy researchers that a moderating effect is identified by looking at interaction effects, e.g., the moderator $\times$ treatment effects (Baron & Kenny, 1986).

The major issue with moderation analysis lies in identifying which moderators should be tested. It is often hard to use theory to select plausible moderators, and a more data-driven approach is then used. Testing many moderators reduces the chance that a moderator will replicate in new samples. However, small sample sizes typically lead to tests with low sensitivity to detect moderators that actually have a clinically meaningful impact on the treatment effect. A further challenge is that it is hard to validate the predictions made using the identified moderators. If we use a patient's age, gender, and education level to predict that they will improve by 10 points more if given PDT versus CBT—how do we know that this prediction is accurate? After the patient has finished

their PDT treatment using their observed outcome does not validate the prediction that the relative improvement compared to CBT is 10 points.

The challenges of building clinical prediction models are much better covered in biomedical literature than in the psychotherapy literature. In medicine, moderators are generally called prescriptive or predictive markers (Janes, Brown, Huang, & Pepe, 2014). Before covering prescriptive models, it is useful to first look at prognostic models, i.e., model that try to predict patients' outcomes (not the effect of the treatment). Best practices for clinical prediction models for prognostic variables are well developed. However, in my experience, psychotherapy researchers tend to ignore the recommendations; mostly by using variable selection strategies that are known to overfit, such as stepwise methods, and report very few metrics on neither the internal or external validity of the model (Bouwmeester et al., 2012; Steyerberg, 2009; Steyerberg et al., 2018). For instance, if we have a prognostic model, we can make a prediction for an individual and compare the prediction to their actual outcome (this agreement between observed outcomes and prediction is called calibration). If we predict that a patient will score ten on an outcome, then it is conceptually straight-forward just to look how close the patient's outcome is to ten. We can summarise such performances using common performance metrics (Steyerberg et al., 2010), e.g., the mean squared error (i.e., R^2 if normalized) or look at calibration curves to see if the model makes accurate predictions for the whole range of outcome values. Generally, a model will perform best on the patients that were used when building the model. It is always noise in the data, and a model that conforms too closely to the original sample, and make accurate predictions for those patients, will most likely perform relatively worse on a new sample. For a model to be useful, it should perform well with new patients. However, in my experience, psychotherapy researchers seldom report estimates of the out-of-sample performance; the performance of a predictor is often reported as significant or non-significant, which says very little about the model's predictive performance. There are statistical methods to help reduce overfitting such as cross-validation and penalized regression, and validation the model on an independent sample (Friedman, Hastie, & Tibshirani, 2001; Harrell, 2015; Steyerberg, 2009).

Prognostic models are important and can be useful when deciding if a patient needs treatment; however, they do not tell us which patients are likely to benefit from which treatment. What we want is a model that can predict if a patient is more or less likely to benefit from, say, PDT treatment or CBT treatment. The challenges that apply to building a prognostic model also applies to predicting treatment effect modification—except that fundamental problem of causal inference makes it much harder to validate the model (Fine & Pencina, 2015; Janes et al., 2014; Janes, Pepe, McShane, Sargent, & Heagerty, 2015; Kent, Steyerberg, & Klaveren, 2018), and the variable selection methods used

with prognostic models tend to fail (Gunter, Zhu, & Murphy, 2007; Lu, Zhang, & Zeng, 2013). For instance, we can build a model that predicts that a patient is expected to improve by 10 points more if given PDT compared CBT. However, we cannot assess calibration by comparing the prediction to the observed outcome—as the individual treatment effects are fundamentally unobservable. Thus, we cannot evaluate and validate our prescriptive model using the same methods and metrics that are generally used with clinical prediction models. Still, in order to build a useful treatment selection model, multivariable prediction models need to be built. We could see if the selected moderators are stable and replicate in a new sample. However, even if, for instance, education and gender, are found to be robust moderators, the treatment effects could vary substantially within these groups. For instance, based on my education and gender, a model might predict that I will benefit by 10 points if I enter treatment A. However, it is not easy to answer the question how accurate such a prediction is for all patients with education = "university or higher" and gender = "male". There are recent attempts to overcome these problems by combining machine learning and causal inference methods (Lipkovich, Dmitrienko, & D'Agostino, 2017; Luedtke & van der Laan, 2017; van Klaveren, Steyerberg, Serruys, & Kent, 2018). However, careful attention should be paid to the unverifiable assumptions that these models need to make—and if it is likely that they hold for psychotherapy treatment selection.

From a practical point-of-view, evaluating the potential clinical usefulness of the decisions made using the predictive model are probably best evaluated in a new RCT where it can be compared to another selection strategy, e.g., if the psychotherapist and the patient selects a treatment without any aid of the new prediction model (Kent et al., 2018). Some would also argue that we should include therapist selection in this decision procedure.

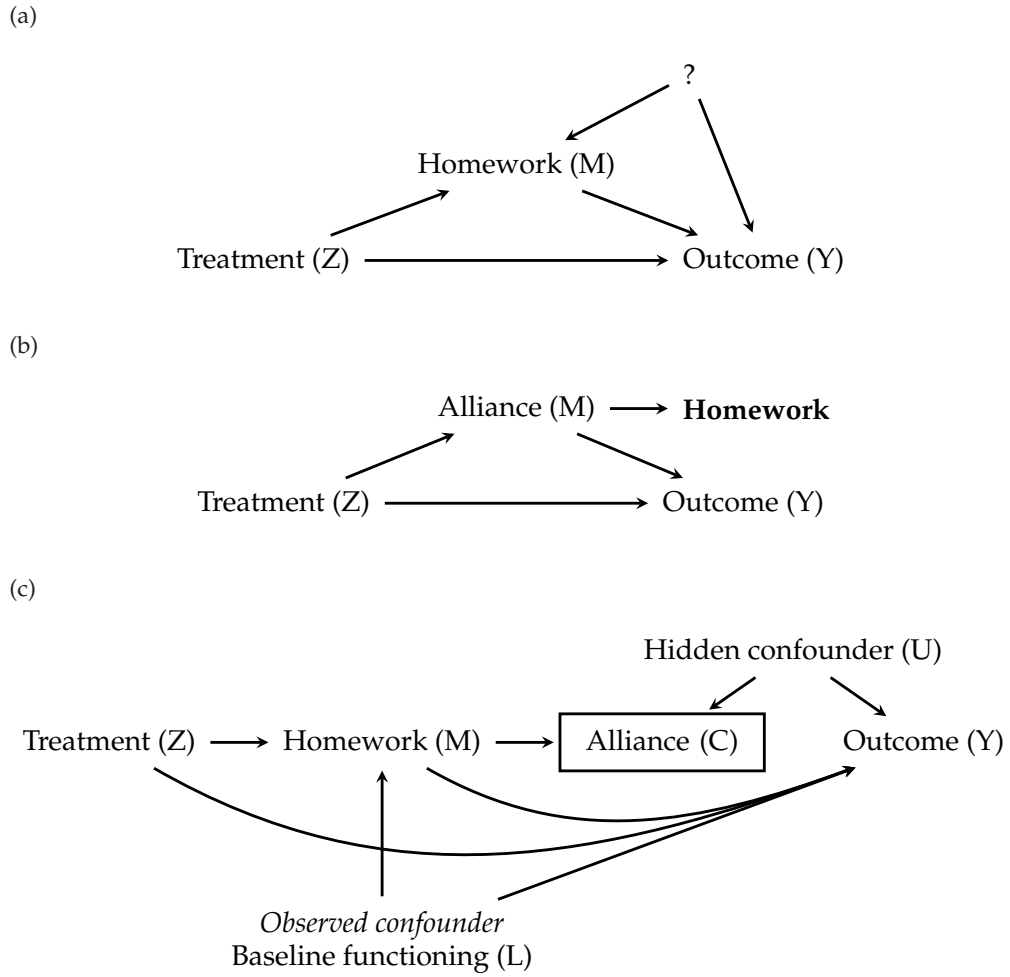


Figure 2.2: A causal diagram of how homework or alliance mediates a treatment effect, where Z represent treatment assignment. a) Shows the typical mediation diagram except that (?) represents a known (L) or an unknown (U) common cause of both homework and the outcome. (b) Shows a simplified scenario were alliance mediates the treatment outcome, and alliance causes homework adherence, while homework has no effect on the outcome. (c) Shows a scenario where homework mediates some of the effect, and where homework adherence influences allegiance and an unknown variable influences both allegiance and the outcome.

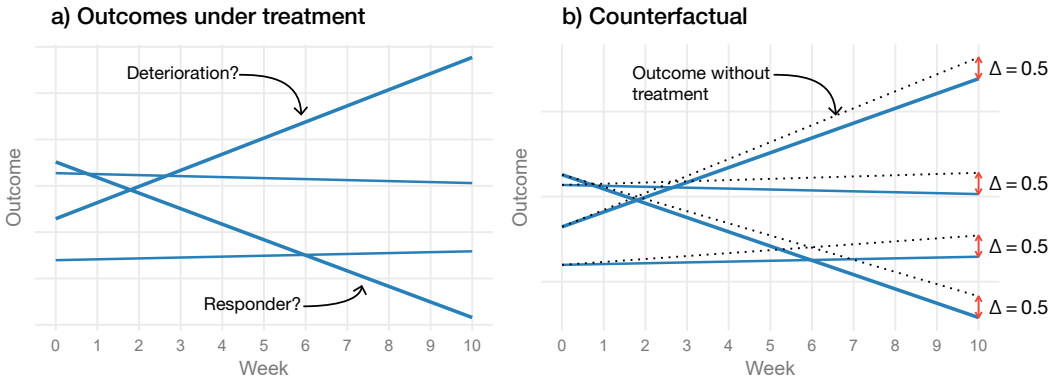


Figure 2.3: The problem of inferring treatment effects on the individual level. (a) Shows the observed change during the treatment, and (b) includes the counterfactual slopes that would have been observed had the participants not recieved treatment. The figure is a longitudinal adaption of Figure 1 in Senn (2016).

Chapter 3

Therapist Effects

A major challenge when evaluating and designing psychological interventions is the possibility that some therapists consistently perform better than others (Lambert, 2013; Norcross, Beutler, & Levant, 2006), which introduces a correlation among subjects belonging to the same therapist. This has been extensively covered for cross-sectional analysis (Baldwin et al., 2011; Wampold & Serlin, 2000). However, with the increased popularity of longitudinal analyses using linear mixed-effects models, these old issues return in a new light.

Among methodologists and statisticians, there is a broad consensus that clustering due to therapists need to be accounted for, and that ignoring a level of nesting will increase the risk of committing a type I error. This was pointed out by Meehl already back in 1955, and Martindale back in 1978, who called it the "therapist-as-fixed-effect fallacy". Moreover, Kiesler (1966) also pointed out the unreasonable assumption of uniform therapist outcomes. Not surprisingly, the reporting of clustering effects is part of the CONSORT statement for non-pharmacological treatments (Grant et al., 2018). Unfortunately, reviews show that investigators mostly do not report therapist effects, and worse, do not adjust their statistical analyses. Martindale (1978) reviewed 33 psychotherapy studies and concluded that few accounted for clustering. Crits-christoph & Mintz (1991) reviewed 140 articles in the *Journal of Consulting and Clinical Psychology* between 1980 and 1990 and found that two-thirds of the studies completely ignored the therapist factor. The journal *Psychotherapy Research* dedicated a whole special issue to therapist effects (Hill, 2006). A more recent study also points out the problems with this type of treatment-related clustering, and that it still continues to be ignored (Walwyn & Roberts, 2015). Similarly, when these individual RCTs are included in a meta-analysis, the problem of therapist effect also applies to the meta-analysis, leading to type I errors that are higher than the specified α-level (Owen, Drinane, Idigo, & Valentine, 2015;

Walwyn & Roberts, 2015).

3.1 How Much Variance is Accounted for by Therapists?

Most studies that have investigated therapist effects find that only a small proportion of the variance in the outcomes after treatment is attributable to the therapists. For efficacy studies, most estimates range from 5 to 10%, with larger numbers being found in naturalistic studies (Baldwin et al., 2011; Kim, Wampold, & Bolt, 2006). In a simple cross-sectional analysis, the proportion of variance at the therapist level is also the intra-class correlation (ICC), which gives the correlation between any two randomly picked patients belonging to the same therapist (Raudenbush & Bryk, 2002).

3.2 The Design Effect

Small ICCs does not mean that the impact on the type I errors is small. The effect of ignoring this multilevel hierarchy on a parameter's standard error (SE) is often called the design effect (Snijders, 2005), which is defined as:

$$D_{eff} = \frac{SE_{correct}}{SE_{incorrect}}.$$

Which tells us by how much we need to multiply the incorrect standard errors to get the correct standard errors. For a cross-sectional analysis with therapists nested within treatments, the design effect is:

$$D_{eff} = \sqrt{(1 + (\bar{n_2} - 1)\rho_1}.$$

Where ρ_1 is the intra-class correlation (ICC) at the therapist level, and $\bar{n_2}$ is the average number of subjects per therapist.

3.3 Therapist or Treatment?

Many authors have discussed the problems of disentangling therapist effects from treatment effects (Chambless & Hollon, 2012; Elkin, 1999). Walwyn & Roberts (2010) discussed the threats of therapist effects to a study's internal and external validity. They noted that when therapists are not randomized to treatments, it is possible that therapist characteristics differ between treatments. In a sense, treatment and therapist effects are confounded. What is being evaluated is the treatment "package" (Elkin, Parloff, Hadley, & Autry, 1985), i.e., both the treatment approach as well as the types of therapists that prefer one treatment orientation. Similar problems have been discussed in medicine, for

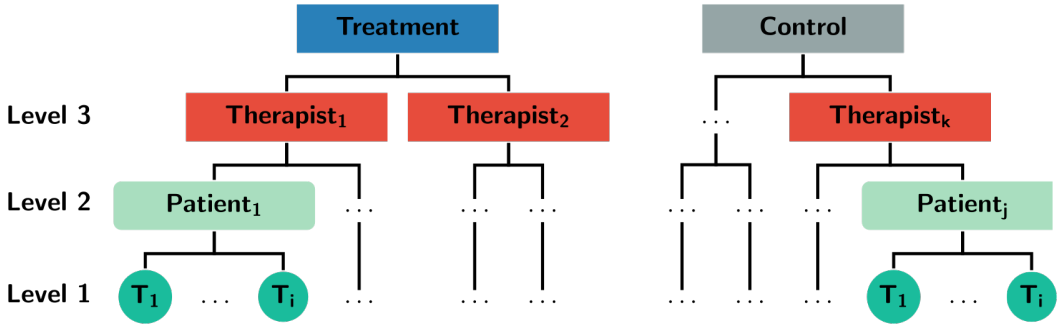


Figure 3.1: An example of a three-level hierarchy where therapists are nested within treatments.

example, when surgeons are nested within treatments. Devereaux et al. (2005) called this an expertise-based trial. Moreover, it is also possible that therapist effects vary between treatments, which could be caused by some treatments being harder to learn, or that less standardized treatments increase the variance between therapists.

3.4 Nested Versus Crossed

If therapists deliver both of the two treatments being compared, then it is possible to disentangle some of the therapist effects from the treatment effects (Walwyn & Roberts, 2010; Wampold & Serlin, 2000). If therapists only deliver one of the treatments, we have a nested design, as shown in Figure 3.1. If therapists deliver both treatments, we have a crossed design. Both designs have their problems and merits. In a nested design, it is not possible to rule out selection bias, therapists nested in Treatment A might overall be more skilled, thus confounding the treatment effect. It also not possible to separate the main effect of therapists, from the possible therapist $\times$ treatment-interaction (Schielzeth & Nakagawa, 2013). Besides making it harder to interpret the therapist effects, this also leads to a loss of power compared to the crossed design (de Jong, Moerbeek, & van der Leeden, 2010). However, in a crossed design, it gets increasingly hard to train therapists in two different therapies simultaneously, and the risk of contamination is evident (Chambless & Hollon, 2012). It is also unlikely that therapists can carry out both treatments with equal skill and commitment, thus making the nested design more realistic in psychotherapy research (Falkenström, Markowitz, Jonker, Philips, & Holmqvist, 2013; Wampold & Serlin, 2000).

3.5 Issues Related to Testing if the Therapist Variance is Exactly Zero

Most studies in clinical psychology are severely underpowered to detect therapist variance. Moreover, testing the statistical hypothesis that the therapist variance is exactly zero, is problematic since we are testing a parameter on the boundary of the sample space. Self & Liang (1987) called this nonstandard testing, and the consequences of testing parameters on the boundary were explored more fully by Stram & Lee (1994), and Stern & Welsh (2000). In short, the standard likelihood ratio test breaks down, making the test too conservative, and thus, it approximately doubles the type II error rate. Concerns regarding type II errors have led some psychologists to proposed a two-step analysis using liberal α-levels of 0.2 to 0.3 (Crits-christoph & Mintz, 1991; Kirk, 2013). However, failing to reject the null hypothesis of no therapist variance does not show support for the parameter being zero. This is a common misunderstanding of significance testing (Nickerson, 2000). The general consensus is that therapists are part of the study design and that the statistical analysis should be congenial with the study design, i.e., clustering should be accounted for even when it is "non-significant" (Baldwin et al., 2011; Roberts & Roberts, 2005).

3.6 Random or Fixed Effects?

Many authors have discussed whether therapists should be viewed as a random or fixed effect (Crits-Christoph, Tu, & Gallop, 2003; Siemer & Joormann, 2003; Wampold & Serlin, 2000). The discussion has mainly revolved around generalizability; fixed effects concerns the therapists in the study, whereas random effects treat the therapists as a sample from a larger population. Proponents of fixed effects often refer to the fact that therapists seldom are a random sample from a larger population (Siemer & Joormann, 2003). Whereas authors arguing for treating therapists as a random factor, point out that we want to make inference about a broader population (Crits-Christoph et al., 2003). This view has been criticized, by the argument that generalizing to a broader population should be based on external validity, and not on a specific statistical model (Kahan & Morris, 2013). It has also been argued that "random sampling" is of marginal importance, "as a (large) number of draws from any cross-section will most likely appear random" (Dieleman & Templin, 2014, p. 4). It should be noted, that under the fixed effects model, statistical inference is made conditional on the included therapists not changing. Which means that from a frequentist perspective, the hypothetical "repeated sampling" would be performed with identical therapists. Whereas, under the random effects model, repeated sampling involves sampling new therapists from the estimated

distribution of therapists. Thus, the assumed data-generating process (DGP) and the inference is substantially different.

Furthermore, Andrew Gellman called the terms "fixed" and "random" effects ill-defined (Gelman, 2005), and prefers to use the term varying instead of random (Gelman & Hill, 2006). In the context of multilevel analysis, Gelman instead differentiates between, complete pooling, no pooling, and partial pooling. Complete pooling ignores all therapist clustering, thus explicitly assuming that all therapists have precisely the same overall success. No pooling corresponds to a fixed effects analysis, which assumes that therapists differ overall in their outcomes and that the outcome from one therapist tells us absolutely nothing about another therapist, i.e., no information is shared between clusters. Partial pooling, which is what standard multilevel/linear mixed-effects models do, assumes that therapists have different outcomes, but that we can share some information between therapists. Partial pooling "borrows information" across therapists, by shrinking estimates closer to the overall mean.

3.7 Therapist Effects and Longitudinal Analyses

Despite the vast number of articles published regarding therapist effects, few have focused on therapist effects in longitudinal designs. Moreover, the ICC, design effect, and power function gets more complicated when therapists are allowed to have varying slopes over time and generally depends on several nuisance parameters (Hedeker & Gibbons, 2006). Thus, recommendations from methodological texts based on cross-sectional data do not necessarily hold. Although, de Jong et al. (2010) considered power for a three-level model with therapists, patients, and repeated measures, they only briefly focused on random slopes at the therapist level. More importantly, they did not investigate the consequences of ignoring therapist effects.

Chapter 4

Gambling Disorder and Semicontinuous Data

> *"If the gambling establishment cannot persuade a patron to turn over money with no return, it may achieve the same effect by returning part of the patron's money on a variable-ratio schedule."*
>
> – B. F. Skinner, *Science and Human Behavior* (1953)

The methodological issues investigated in this thesis are applied to two studies on gambling disorder. In one study, we investigate the agreement between a collateral and the person that gambles on the amount of gambling losses, and in the other we perform an RCT to evaluate an internet-delivered support program for the concerned significant others of problem gamblers. Following is an overview of both gambling disorder and interventions aimed at CSOs, which is followed by an introduction to the problem of analyzing gambling expenditures as a treatment outcome; a problem dealt with in Study II and Study III.

4.1 Gambling Disorder

Gambling disorder is generally categorized into problem gambling and pathological gambling. While pathological gambling is defined in both DSM-5 and ICD-11, problem gambling is a broader term used to describe a less severe form of gambling problems. Problem gambling is characterized by the inability to control time spent and money wagered on gambling, despite having a negative impact on the gambler's economy, emotional well-being, and social relations. Problem gambling is often associated with trying to win back money lost on gambling, using gambling to cope with depressive feelings, having to loan money to pay expenses or to gamble more, or lying about the time and money spent on gambling (Hodgins, Stea, & Grant, 2011; Petry & Weiss, 2009).

About 2% of the Swedish population between 16 and 85 years of age experience problems caused by gambling (Abbott, Romild, & Volberg, 2018). Prevalence estimates for gambling problems vary across countries, from 0.2% in Norway to 5.3% in Hong Kong (Hodgins et al., 2011). This is generally attributed to differences in accessibility and availability of gambling, but also to differences in survey methods such as screening techniques, timeframe, administration, and response rates (Hodgins et al., 2011). However, results from different prevalence studies indicate that gambling problems in Sweden seems to be relatively stable across time (Abbott et al., 2018; Volberg, Abbott, Rönnberg, & Munck, 2001). Gambling problems are unevenly distributed in Sweden, with a higher prevalence among men, people born outside of Sweden, people with a low eduction, and people 18 to 24 years old (Abbott, Romild, & Volberg, 2014).

Problem gambling is associated with psychological distress (Barry, Stefanovics, Desai, & Potenza, 2011). Psychiatric comorbidity is common, especially substance-related disorder, as well as anxiety and affective disorders (Håkansson, Karlsson, & Widinghoff, 2018; Håkansson, Mårdhed, & Zaar, 2017; Williams, Volberg, & Stevens, 2012). Furthermore, suicide attempts and suicide mortality are more common among people with gambling problems (Karlsson & Håkansson, 2018; Newman & Thompson, 2007).

Psychological interventions, such as cognitive-behavioral therapy is the most well-supported treatment for problem gambling (Cowlishaw et al., 2012). However, evaluating gambling treatments is challenging. Many studies of gambling treatments are faced with a substantial relapse rate (Echeburúa, Fernández-Montalvo, & Báez, 2001), and high attrition rates (Westphal, 2007), severely affecting what conclusions are possible to draw. Moreover, there seems to be large non-specific effects associated with just deciding to seek treatment. Placebo, or non-specific response to gambling treatment, has been discussed by several authors (Carlbring, Jonsson, Josephson, & Forsberg, 2010; Toneatto & Ladoceur, 2003; Westphal, 2008). Some empirical evidence for non-specific effects exists in the literature. For instance, it has been found that problem gamblers respond to very minimal interventions, such as reading a 30-page booklet on CBT, receiving one session of motivational interviewing, receiving just a clinical interview or even being put on a waiting list (Diskin & Hodgins, 2009; Hodgins, Currie, & el-Guebaly, 2001; Petry, Weinstock, Ledgerwood, & Morasco, 2008). These non-specific effects, combined with high attrition rates, make it hard to draw conclusions about the long-term effects of gambling treatments.

4.1.1 Concerned Significant Others

Gambling can not only be devastating for the gambler, it can also have serious negative effects on the lives of the concerned significant others (CSOs; Langham et al., 2016). A

large portion (18%) of the adult Swedish population sees themselves as CSOs of problem gamblers (Svensson, Romild, & Shepherdson, 2013). CSOs of problem gamblers tend to report worse physical and psychological health, and that the relationship to the gambler is harmed (Kalischuk, Nowatzki, Cardwell, Klein, & Solowoniuk, 2006; Volberg et al., 2001). In a representative sample in Norway, Wenzel, Øren, & Bakken (2008) found that 63% of the CSOs reported that the gambler had worsened the family's financial situation, and 65% reported that the gambling had led to conflicts in the family. Many CSOs report that they are often left feeling isolated and unsupported (Krishnan & Orford, 2002). However, CSOs of gamblers can play an essential role in recovery. For instance, as many as 50% of problem gamblers report that they rely on informal help provided by their CSO (Clarke, Abbott, DeSouza, & Bellringer, 2007), and gamblers report concerns for CSOs as an important reason for entering treatment (Bertrand, Dufour, Wright, & Lasnier, 2008).

There is evidence that shame and stigma are the main barriers for CSOs in seeking help (Hing, Tiyce, Holdsworth, & Nuske, 2013; Valentine & Hughes, 2010), and that CSOs typically turn to self-help, online or telephone support before seeking professional help (Hing et al., 2013). Thus, it is possible that an internet-delivered treatment could seem attractive to CSOs.

Despite the long list of gambling-related negative consequences that CSOs suffer, support for CSOs has been limited (Hodgins, Toneatto, Makarchuk, Skinner, & Vincent, 2007). A mere handful of studies have evaluated interventions for CSOs of problem gamblers. The types of interventions available for CSOs of addicts can broadly be categorized into three categories: 1) working with the CSO to motivate the addict to enter treatment, 2) involving a CSO in the treatment of the addict, and 3) working with the CSO's needs in their own right (Copello, Velleman, & Templeton, 2005).

4.1.2 CSOs and Problem Gamblers' Motivation to Seek Treatment

Only about 5% of the problem gamblers seek professional help (Cunningham, 2005; Statens folkhälsoinstitut, 2010). Numerous researchers have suggested that CSOs can play a crucial role in getting the gambler to enter treatment, and they have highlighted the need to equip CSOs better to handle the problem gambling (Clarke et al., 2007; Dickson-Swift, James, & Kippen, 2005; Gomes & Pascual-Leone, 2009; Ingle, Marotta, McMillan, & Wisdom, 2008; Petry & Weiss, 2009). Even though financial concerns are often the main reason that gamblers seek help (Bellringer, Pulford, Abbott, DeSouza, & Clarke, 2008), many gamblers report concerns for CSOs as an important reason for entering treatment (Hing et al., 2013; Hodgins & el-Guebaly, 2000).

4.1.3 Community Reinforcement and Family Training

Research on training-programs aimed at CSOs of substance misusers has shown promising results in getting treatment-refusing substance misusers into treatment. The approach with the most substantial empirical support is community reinforcement and family training (CRAFT; Copello et al., 2005; Fernandez, Begley, & Marlatt, 2006; Meis et al., 2013). The CRAFT-model is based on cognitive-behavioral therapy (CBT)-principles and has three main goals: 1) motivate to the substance misuser to seek treatment, 2) decrease the substance misuse, and 3) increase the CSOs quality of life. A meta-analysis of CRAFT-studies found that, overall, 66% of the CSOs managed to get their loved ones to enter treatment (Roozen, Waart, & Kroft, 2010), whereas the corresponding numbers were 18% for Al/Nar-anon and 30% for Johnson Intervention. Also, CRAFT was found to improve CSO functioning in terms of depression, anger, family conflicts, and relationship happiness.

The CRAFT-model is founded on principles from CBT, especially operant conditioning. The CRAFT-method employs six overall concepts: 1) functional analysis of the substance misuse, 2) communication training, 3) positive reinforcement of sober behavior, 4) the use of natural negative consequences, 5) helping the CSO enrich their own lives, and 6) teaching the CSO when and how to invite the substance misuser to enter treatment.

4.1.4 CRAFT and Problem Gambling

The CRAFT approach has been modified and tested with CSOs of problem gamblers in three studies. Makarchuk, Hodgins, & Peden (2002) first evaluated CRAFT for gambling in a pilot RCT, where a 45-page self-help manual was developed and evaluated. The study compared a group that received the CRAFT-manual to a control condition that received a standard information packet. Both groups displayed significant improvements, but there was no difference in treatment engagement. However, the CRAFT-group reported a greater reduction in gambling, a greater amount of satisfaction with the program, and having their needs met to a larger extent than the control condition. The same research group proceeded with evaluating the CRAFT-program in a larger RCT (n = 186; Hodgins et al., 2007). In this study, they added a CRAFT-condition that received minimal telephone support (1 to 2 calls). Unfortunately, the second study yielded essentially the same results as the pilot study—i.e., inconclusive results regarding any difference in treatment entry between the groups, but a significant difference in favor of CRAFT on days gambling, CSOs' program satisfaction and experiences of having their needs met. The authors concluded that the approach was promising, but that it is likely that CSOs are in need of additional support in order to successfully implement the CRAFT-techniques. Nayoski & Hodgins

(2016), therefore, tested the CRAFT approach in a small study (n = 32), where they compared a face-to-face treatment to a workbook-only group. No conclusive results where found, but effect sizes were in the same direction as previous studies. CSOs in the individual treatment reported greater functioning, and that their gambler spent less money and time on gambling compared to the workbook-only group.

4.1.5 Working with the CSOs in Their Own Right

Few studies have evaluated interventions that focus on working with CSOs of problem gamblers in their own right. My review of the literature only identified one such study. Rychtarik & McGillicuddy (2006) performed a preliminary evaluation of a coping skills training (CST) program for CSOs of pathological gamblers. They found a large reduction in depression and anxiety in the CST-group relative to a wait-list control. However, they could draw no conclusions regarding differences between the groups on partner gambling or treatment entry.

4.2 Semicontinuous Gambling Data

"All models are wrong but some are useful."

– George E. P. Box, *1979*

An important aim of gambling treatments is to reduce gambling losses and help prevent relapses. In a consensus statement regarding the reporting of outcomes from problem gambling trials, it was proposed that measures of gambling behavior should focus on net expenditure and days gambled (Walker et al., 2006). In the published literature, these outcomes are often analyzed as if they were normally distributed, or by log transforming the outcome. Net expenditure on gambling is typically heavily skewed with some participants losing much more money than the rest. Adding further complexity, many participants stop gambling or gamble on very few days when they enter a treatment trial, resulting in data with a lot of zeros (no expenditure). Data on the daily losses on gambling, or the daily number of drinks has typically been collected by retrospective reports, such as the timeline follow-back (TLFB) method. However, the internet and smartphones have made electronic collection much more feasible. More intensive data collection methods, such as diary methods, or ecological momentary assessment, are gaining in popularity. Moreover, gambling research might be unique in the possibilities offered by behavioral tracking of online gambling. By collaborating with gambling operators, researchers get access to ecologically valid data on a transactional level. These research opportunities will likely increase; for instance, an increasing

amount of gambling in Sweden is performed online (Folkhälsomyndigheten, 2019). Behavioral tracking is also increasing; all gambling on Svenska spel's land-based and online products are tracked. There will be a need to evaluate the responsible gambling tools offered by the operators, and the data generated will require sophisticated statistical methods to gain insight into the gambling behavior of the consumers.

4.2.1 Similar Problems in Other Research Fields

Other addiction sub-fields face similar issues. Studies that collect drinks per day or cigarettes per day also include a lot of zeros (Atkins, Baldwin, Zheng, Gallop, & Neighbors, 2013; Bandyopadhyay, DeSantis, Korte, & Brady, 2011). To analyze these data methodologists have mostly proposed different count models, most commonly zero-inflated or hurdle models. Both zero-inflated and hurdle models split the model into two parts: a Bernoulli part for modeling abstinence, and a count distribution (e.g., Poisson or negative binomial) for the number of drinks on drinking days. In the case of hurdle models, the count process is truncated at zero, whereas zero-inflated models allow both parts of the model to contribute zeros. The two types of zeros in a zero-inflated model are often called "structural" and "sampling" zeros (He, Tang, Wang, & Crits-Christoph, 2014). In addiction research, individuals that are not at risk of using, e.g., non-smokers, are viewed as structural zeros, whereas smokers that happened to not smoke during the sampling period are sampling zeros, i.e., they are at risk but did not smoke. In a hurdle model, only the Bernoulli part contribute zeros, so all individuals are conceptualized as being at-risk; thus, zeros are sampling zeros. Hurdle models are also described by a two-step decision process, where individuals first decide if they should drink or not, and once this "hurdle" is crossed a user then decides on how much to drink. Participants in clinical trials are often defined as at-risk simply due to the study's inclusion criteria, and the hurdle model is often preferred. However, a zero-inflated model might still be useful if there are reasons to suspect two distinct processes that generate zero-observations.

DeSantis et al. (2013) found that a hurdle-Poisson model worked well to evaluate treatment effects from high-resolution drinking data. They also found that placing the hurdle at a "low-risk"-cutoff of 4 to 5 for the number of drinks per day, fit the data better than a hurdle at zero. Xing et al. (2015) proposed a two-part Bayesian random-effects model with a skewed distribution to model dependency symptoms data. Atkins et al. (2013) proposed the hurdle Poisson model for count drinking data in treatment research, and Bandyopadhyay et al. (2011) also found that a hurdle binomial model best fit their data. However, these are mostly count models, and their findings are unlikely to generalize to net losses from gambling.

Continuous dependent variables with excess zeros have a long history in the econometrical literature. Similar to the count models described earlier, the data-generating process is characterized by a two-part economic decision process. One process governs participation (binary part), and another governs the amount of money to spend (continuous part). For instance, studies on health service use typically contain a non-trivial proportion of individuals that did not use health services, and thus did not have any health expenditure during the study period. In these settings, commonly used models are the Tobit, Heckman sample selection model, and two-part model (Basu & Manning, 2009; Mihaylova, Briggs, O'Hagan, & Thompson, 2011). The main difference between the models is the assumption about how zeros arise (Neelon, O'Malley, & Smith, 2016). In the Tobit model, zeros are censored normal observations, and predictors are assumed to have the same influence on the decision to participate and the intensity. Whereas, both Heckman and two-part models separate the decision to participate from how much to spend (Wooldridge, 2010). Conceptually, zeros in the Heckman model represents censored positive values, that could have been observed under ideal circumstances, whereas the two-part model treats zeros as actual zero expenditures. Duan, Manning, Morris, & Newhouse (1983) noted that when zeros represent actual zero expenditure, two-part models are easier to interpret. Moreover, two-part models are more numerically stable (Min & Agresti, 2002). Selection models can be used to model actual outcomes, but are typically used when zeros represent missing values, hence the name sample selection models (Madden, 2008). Additionally, as an alternative to two-part models, Deb & Trivedi (2002) proposed finite-mixture models, to capture e.g., "frequent" and "infrequent" use.

These models have been discussed in gambling research, mostly related to lottery participation, e.g., Humphreys, Lee, & Soebbing (2010) used them to study consumer behavior in lotteries and found that the hurdle model best fit their data. Similar results were found by Rude, Surry, & Kron (2014), who studied Swedish gambling expenditure, and Jaunky & Ramchurn (2014) when modeling consumer behaviors on scratch card markets. Economic models of gambling expenditure have mostly used Tobit or two-part models (Abdel-Ghany & Sharpe, 2001; Crowley, Eakins, & Jordan, 2012; Farrell & Walker, 1999; Sawkins & Dickie, 2002), possibly favoring two-part models. For instance, Stranahan & Borg (1998) made the point that the decision to gamble should be statistically separated from the decision on how much to spend gambling. However, all these are cross-sectional models applied to non-clinical data, mostly applied to population expenditure and participation in lotteries or scratch cards.

4.2.2 Longitudinal Extensions

Two-part models have been extended to longitudinal analyses by incorporating random effects into each part (Olsen & Schafer, 2001; Tooze, Grunwald, & Jones, 2002). In clinical research, it is highly likely that the two parts of the model are correlated, e.g., it is possible that individuals that are more likely to participate in gambling, also are likely to gamble for more money. With longitudinal data it is also possible to have correlated random slopes between the two parts of the model, meaning that change in the probability of participation over time, is correlated with the change in expenditure over time. Tooze et al. (2002) described how this could be achieved by letting the two parts of the model be correlated via their random effects. Independence between the two parts are often assumed due to computational reasons; however, Su, Tom, & Farewell (2009) have shown that ignoring the correlation will bias the results.

4.2.3 Appropriate Treatment Effect Estimands

An issue with two-part models is that they lead to two treatment effects, one for each part of the model—one effect on reporting a zero, and one effect on the impact of the treatment on the non-zero values. It is possible to average over the two parts to get a "marginal" treatment effect; however, due to the nonlinear transformations, this marginal effect will be heterogeneous over the random effects (Smith, Neelon, Preisser, & Maciejewski, 2015). This property of two-part random effects models seems to have been largely overlooked in the addiction treatment studies that use them. However, instead of modeling expenditure conditional on it being non-zero, it is possible to directly model the overall expenditure in the continuous part of the model, by solving for the marginal mean (Smith et al., 2015). These marginalized two-part random effects models will lead to treatment effects on the overall expenditure that are homogenous over the random effects. Yet, as noted by Zhang, Liu, & Hu (2018), these models are not truly marginal models, in the sense that they estimate population-average estimates, but instead estimate treatment effects that are conditional on subject-specific random effects (Diggle et al., 2002). However, this conditional property of the model applies to all (generalized) linear mixed-effects models, and it is not necessarily something negative. Choosing between population-average or subject-specific effects depends on the research question. However, subject-specific models are probably most useful for clinical research, and population-average in public health research (Fitzmaurice, Laird, & Ware, 2012).

Chapter 5

Aims

The aim of this thesis was to explore methodological challenges that impact the evaluation of psychological treatments in general but also gambling treatment trials specifically. The first two studies discuss important methodological issues that researchers tend to overlook, and the issues are investigated empirically using Monte Carlo methods. The results from Study I and II are then applied to answer the clinical research questions in Study III and IV. Specifically, **Study I** investigates the consequences of ignoring therapist effects in longitudinal data. **Study II** investigates the challenges of estimating treatment effects in gambling studies using gambling expenditure as an outcome. **Study III** applies and extends the work in Study II to investigate how concordant gamblers and their concerned significant others are in their reports of gambling losses. **Study IV** applies the results from Study I, II, and III to investigate the effects of an internet-delivered program aimed at the CSOs of treatment refusing gamblers.

Chapter 6

Empirical Studies

6.1 Details on the Methods Used

Before covering the individual studies in this thesis, I will first provide some further elaborations on the key methods used in these studies.

6.1.1 Monte Carlo Simulation Studies

Paper I and II use computer simulations to run experiments in order to evaluate statistical methods empirically. In short, a Monte Carlo study is performed by coding the DGP and drawing samples from it using a pseudo-random number generator. This creates simulated data sets from a known process. We can then use this to evaluate the performance of statistical methods, both the performance of a correct model and a misspecified model that deviates from the true DGP in some important way. For instance, if we generate 5000 data sets and fit a model to each data set, then we can check how many of the 5000 95% confidence intervals that include the true value. If the confidence intervals are valid, 95% ($\pm$ Monte Carlo error) of them should include the true value. Monte Carlo studies are empirical experiments, and substantive knowledge is needed to design useful simulation experiments (Burton, Altman, Royston, & Holder, 2006; Morris, White, & Crowther, 2019).

6.1.2 Power Analysis

Sample size planning for two- and three-level LMMs is challenging, in anything but trivial models, multiple interacting factors impact power. For psychotherapy studies, the therapist level is a complicating factor since most studies only include very few therapists. For these designs, power is significantly impacted by the number of

therapists. Moreover, when the study design is partially nested or when the number of subjects per therapist varies, the correct degrees of freedom must be approximated. In this section, I will give some concrete examples of the challenges as applied to Study IV. When the trial was planned power was based on a two-level random intercept and slopes model, and I partly developed the R package `powerlmm` to solve some of the challenges related to incorporating therapist effects and missing data in the power analysis.

The three-level partially nested model can be written as,

Control group:	*Treatment group:*
Level 1	Level 1
$Y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + R_{ij}$	$Y_{ijk} = \beta_{0jk} + \beta_{1jk} t_{ijk} + R_{ijk}$
Level 2	Level 2
$\beta_{0j} = \delta_{000} + U_{0j}$	$\beta_{0jk} = \gamma_{00k} + U_{0jk}$
$\beta_{1j} = \delta_{100} + U_{1j}$	$\beta_{1jk} = \gamma_{10k} + U_{1jk}$
	Level 3
	$\gamma_{00k} = \delta_{000} + \delta_{001} + V_{0k}$
	$\gamma_{10k} = \delta_{100} + \delta_{101} + V_{1k}$

(6.1)

Where δ_{101} gives the difference in change over time between the treatment and the control group. In addition, the patient-specific and therapist-specific random effects follow multivariate normal distributions,

$$\begin{pmatrix} U_{0jk} \\ U_{1jk} \end{pmatrix} \sim \mathcal{N} \left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{matrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{matrix} \right), \tag{6.2}$$

and,

$$\begin{pmatrix} V_{0k} \\ V_{1k} \end{pmatrix} \sim \mathcal{N} \left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{matrix} \sigma^2_{v_0} & \sigma_{v_{01}} \\ \sigma_{v_{01}} & \sigma^2_{v_1} \end{matrix} \right), \tag{6.3}$$

and the within-patient residuals are, $R_{ijk} \sim \mathcal{N}(0,\ \sigma^2_e)$.

Power will depend greatly on the relative size of the variance components. Two important ratios are the "variance ratio", $(\sigma^2_{u1} + \sigma^2_{v1})/\sigma^2_e$, as well as the amount of variance in change over time attributed to the therapist level (the "therapist effect"), $\sigma^2_{v1}/(\sigma^2_{u1} + \sigma^2_{v1})$. The technical details of the calculations are covered in Magnusson (2018), and in `powerlmm`'s vignettes. Figure 6.1 shows the power curves for both the achieved

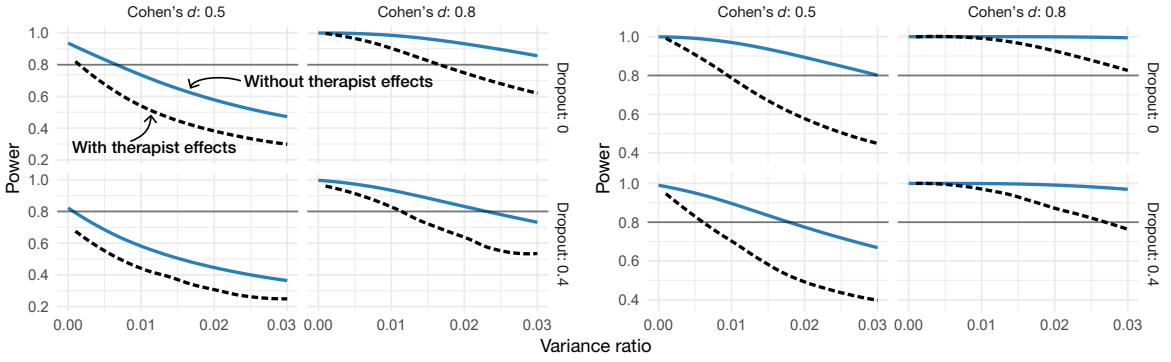


Figure 6.1: Power as a function of the variance ratio for different effect sizes and amounts of dropout (proportion of dropout at the last time point). The calculations are based on the realized sample size (100 participants, unbalanced allocation to therapists) and 7 time points. The dashed curved shows power when 5% of the slope variance is at the therapist level.

and planned for sample size in Study IV. We see that with the achieved sample size the statistical test will have rather low sensitivity assuming that the true effect would be $d = 0.5$, unless the random slope variance is small compared to the within-subject residual variance.

As there is a possibly large difference in power under the assumption of a two-level or three-level model, one might be tempted to use a likelihood ratio test (LRT) to rule out any therapist effects. Figure 6.2 shows that the likelihood ratio test does not automatically keep the type I error at correct levels.[1] There is always a balance between type I errors and power, as the LRT α is increased both power and type I errors moves towards the three-level model that always accounts for therapist effects. Moreover, the comparison is skewed towards the two-level model since we are effectively accepting a larger α level. The difference is even less pronounced if we set $\alpha = 0.075$ for the three-level model. If investigators are willing the increase the risk of committing a type I error in order to reduce the risk of a type II error, then the more principled way of achieving this would be to increase α together with using a three-level model. The problem with using the two-level model is that the actual α level depends on unknown factors, such as the between-therapist variance in change over time. The problem with the three-level model is that power depends a lot on the number of therapists. Thus, from a planning point of view, it might be hard to design studies with a reasonable chance of detecting clinically relevant effects. Figure 6.3 shows that just adding more participants does very little to increase power; the number of therapists will have a much larger impact on power than the total number of subjects.

[1]This example is adapted from one of my blog posts: https://rpsychologist.com/do-you-need-multilevel-powerlmm-0-4-0

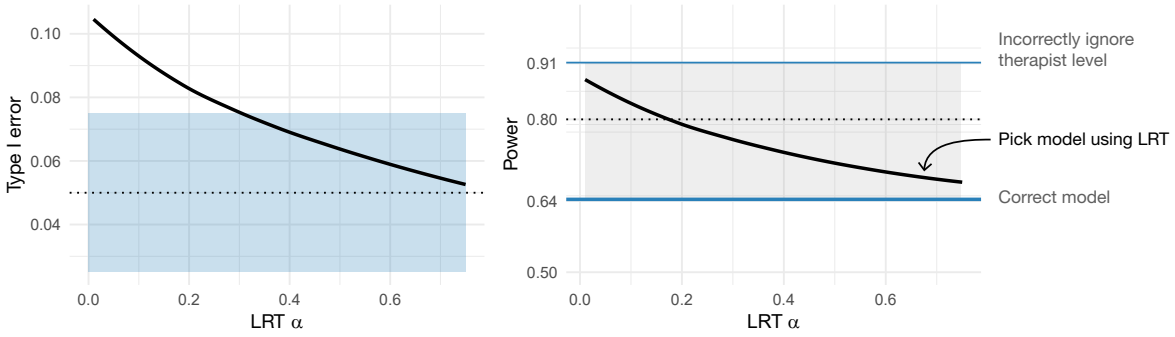


Figure 6.2: The impact of using LRT model selection to decide if therapist effects should be accounted for. Impact of the LRT's α-level is shown on both the type I errors and power.

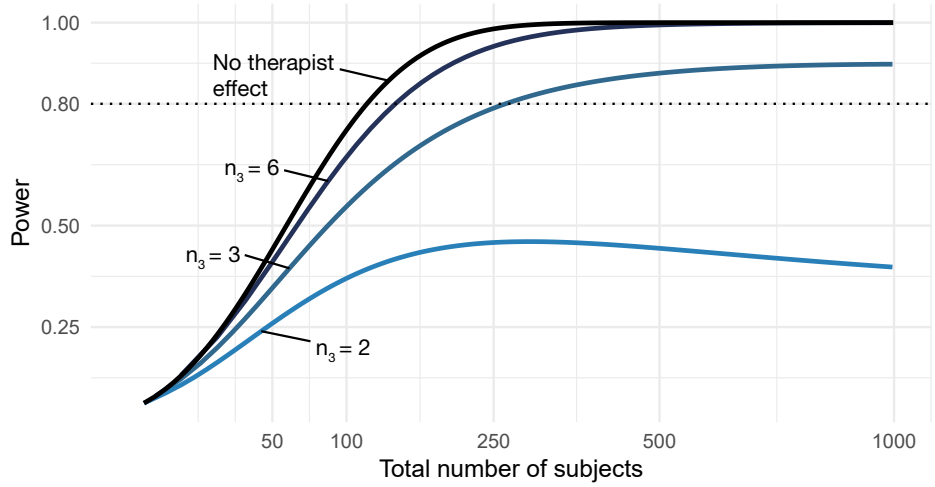


Figure 6.3: The impact of the number of therapists on power as a function of the total sample size. The curve labeled *No therapist effect* shows power assuming a two-level model with no therapist-level random slope variance.

Generally, both dropout and the number of subjects per therapist will be unknown, and we must use some approximation to decide how sensitive we want our test to be given a reasonable level of therapist imbalance and missing data.

6.1.3 Missing Data Considerations

Psychological treatments are delivered over repeated sessions, and patients are generally followed-up for 6 to 12 months after treatment completion—making missing data practically unavoidable. This is probably the main reason for the popularity of longitudinal data analysis via linear mixed-effects models in clinical psychology.

Rubin (1976) presented three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR). LMMs provide unbiased estimates under MAR missingness; however, it is not entirely clear

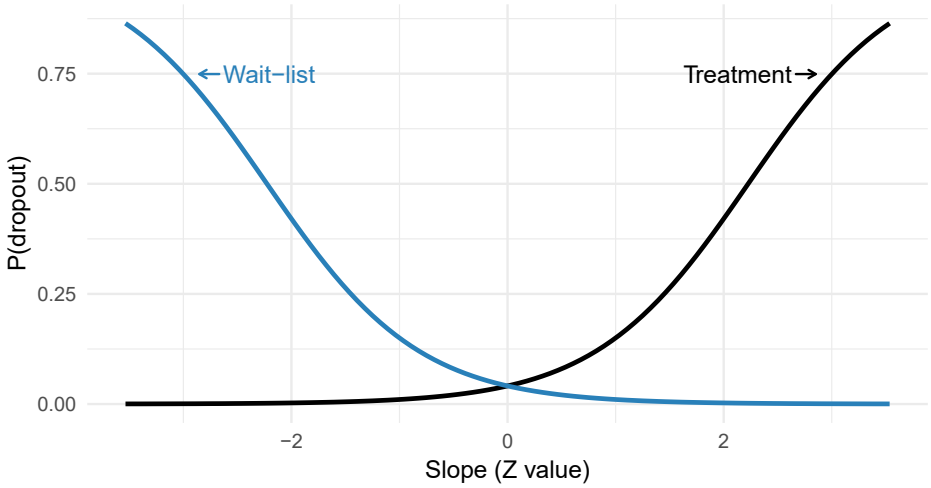


Figure 6.4: A differential MNAR dropout process where the probability of dropping out from a trial depends on the patient-specific slopes which interact with the treatment allocation. The probability of dropout is assumed to be constant over time.

that assuming MAR is completely justified. I will, therefore, present a simple example of when the LMM fails and illustrate some of the sensitivity analyses used in Study IV.

Simplified, if we have the complete outcome variable Y (which is made up of the observed data Y_{obs} and the missing values Y_{miss}) and a missing data indicator R (Little & Rubin, 2014; Rubin, 1976; Schafer & Graham, 2002), then we can write the MCAR and MAR mechanisms as,

$$\begin{aligned} \text{MCAR}: &\quad \mathrm{P}(R \mid Y) = \mathrm{P}(R) \\ \text{MAR}: &\quad \mathrm{P}(R \mid Y) = \mathrm{P}(R \mid Y_{obs}). \end{aligned} \tag{6.4}$$

If the missingness depends on Y_{miss}, the missing values in Y, then the mechanism is MNAR. MCAR and MAR are called ignorable because the precise model describing the missing data process is not needed. In theory, valid inference under MNAR missingness requires specifying a joint distribution for both the data and the missingness mechanisms (Little, 1995). There are no ways to test if the missing data are MAR or MNAR (Molenberghs, Beunckens, Sotto, & Kenward, 2008; Rhoads, 2012), and it is therefore recommended to perform sensitivity analyses using different MNAR mechanisms (Hedeker & Gibbons, 1997; Little, 1995; Schafer & Graham, 2002).

6.1.3.1 An Empirical Example of MAR vs. MNAR Missing Data

LMMs are frequently used by researchers to deal with missing data problems in psychotherapy trials. However, in my opinion, researchers frequently misunderstand the MAR assumption and fail to build a model that would make the assumption more plausible. Sometimes you even see researchers using tests, e.g., Little's MCAR test, to

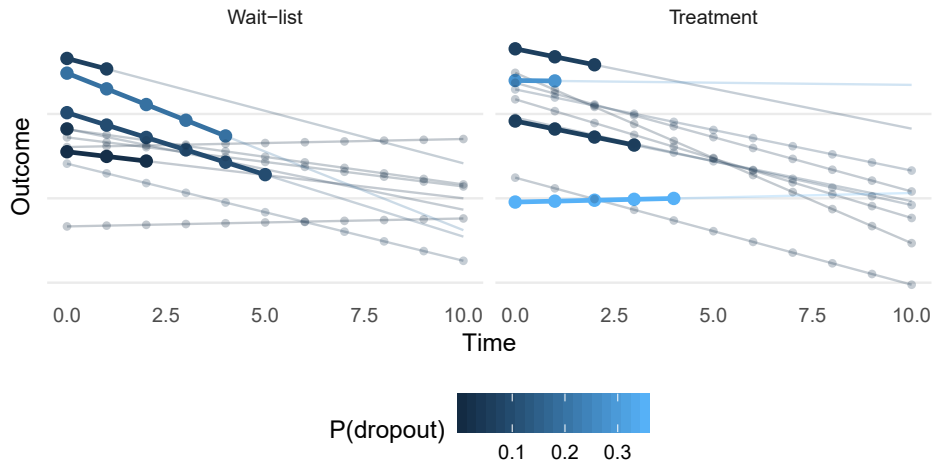


Figure 6.5: A sample of patients drawn from the MNAR (random slope) data-generating process. Circles represent complete observations; the bold line represents the slope before dropping out. P(dropout) gives the probability of dropout, which is assumed to be constant at all time points.

prove that the missing data mechanisms is either MCAR or MAR and hence ignorable. Naturally, as stated earlier, such a conclusion is a clear misunderstanding and builds on faulty logic.

Another common misunderstanding is that LMMs yield unbiased estimates if the dropout is related to the patient-specific slopes (i.e., the random effects). Clearly, it would be practical if the inclusion of random slopes would allow missingness to depend on patients' latent change over time. Unfortunately, the random effects are latent variables and not observed—hence, such a missingness mechanism would also be MNAR (Little, 1995). Figure 6.6 illustrates the MAR, outcome-based MNAR, and random coefficient-based MNAR mechanisms. In my experience, psychotherapy researchers tend to include quite few prognostic covariates in their models, or variables that are thought to be related to the missingness. Thus, most LMMs (that are used to deal with missing data) make the assumption that missingness only depends on the previously observed values of the outcome. This is quite a strong assumption.

To illustrate these concepts I generated data from a two-level LMM with random intercept and slopes, and included a MNAR missing data mechanism where the likelihood of dropping out depended on the patient-specific random slopes. Moreover, the missingess differed between the treatment and control group. Figure 6.4 and 6.5 illustrate the dropout mechanism, which are based on the following equations,

$$\begin{aligned} \text{logit}(\Pr(R_{ij} = 1 | TX_{ij} = 1)) &= -\sigma_{u_1} + \text{logit}(0.15) + U_{1j} \\ \text{logit}(\Pr(R_{ij} = 1 | TX_{ij} = 0)) &= -\sigma_{u_1} + \text{logit}(0.15) - U_{1j}. \end{aligned} \tag{6.5}$$

Table 6.1: Results from the simulation with a random slope MNAR dropout process.

Model	$\overline{\text{Est.}}$	Rel. bias (%)	$\bar{d}$	Power	CI coverage	Type I error
LMM (MAR)	-11.84	274	-0.74	1.00	0.02	0.83
GEE	-11.19	254	-0.70	1.00	0.06	0.71
LMM (PM)	-5.39	70	-0.34	0.64	0.84	0.10
JM	-3.18	0.6	-0.20	0.28	0.93	0.07
LMM (complete)	-3.21	1.4	-0.20	0.38	0.95	0.05

Note. MAR = missing at random; LMM = linear mixed-effects model; GEE = generalized estimating equation; JM = joint model; PM = pattern mixture; $\overline{\text{Est.}}$ = mean of the estimated effects; Rel. bias = relative bias of $\overline{\text{Est.}}$; $\bar{d}$ = mean of the Cohen's d estimates.

To show the consequences of this random-slope dependent MNAR scenario under different models, I performed a small simulation. The study had 11 time points, 150 participants per group, the variance ratio was 0.02, the pretest ICC was 0.6, with a correlation between intercept and slopes of -0.5, and there was a small effect in favor of the treatment of $d = -0.2$. Five different models was compared:

- LMM (MAR): a classical LMM assuming that the dropout was MAR.
- GEE: a generalized estimating equation model.
- LMM (PM): an LMM using a pattern-mixture approach. Two patterns were used; either "dropout" or "completer", and the results were averaged over the two patterns.
- JM: A joint model that correctly allowed the dropout to be related to the random slopes.
- LMM with complete data: an LMM fit to the complete data without any missingness.

Table 6.1 shows the results, and Figure 6.7 shows how much the treatment effects differ. We can see that LMMs are badly biased under this missing data scenario; the treatment effect is much larger then it should be (Cohen's d: -0.7 vs. -0.2). The pattern-mixture approach improves the situation, and the joint model recovers the true effect. Since the sample size is large, the bias under the MAR assumption leads to the LMM's CIs having extremely bad coverage. Moreover, under the assumption of no treatment effect the MAR LMM's type I errors are very high (83%), whereas the pattern-mixture and joint model are closer to the nominal levels.

This simulation example is purposely quite extreme. However, even if the MNAR mechanism would be weaker, the LMM will yield biased estimates of the treatment effect. The assumption that dropout might be related to patients' unobserved slopes is not unreasonable. However, fitting a joint model is often not feasible as we do not

know the true missingness mechanism. I included it just to illustrate what is required to avoid bias under a MNAR mechanism. In reality, the patients' likelihood of dropping out is likely an inseparable mix of various degrees of MCAR, MAR, and MNAR mechanisms. The only sure way of avoiding bias would be to try to acquire data from all participants—and when that fails, perform sensitivity analyses using reasonable assumptions of the missingness mechanisms.

6.2 Study I: "The Consequences of Ignoring Therapist Effects in Longitudinal Data"

In Study I, we investigated the consequences of ignoring therapist effects in longitudinal data analysis. We also performed a small review of all the trials published in the *Journal of Consulting and Clinical Psychology* from 2008 to 2018, which showed that virtually no investigators accounted for the possibility that therapists might differ in their effectiveness.

We present what factors influence type I errors when ignoring therapist effects and also report the results from a large simulation study.

6.2.1 Methods

We first analytically derived what factors would impact the Type I errors when ignoring therapist effects, which were: the number of time points, the number of subjects per therapist, the amount of heterogeneity in change over time, and the amount of the total slope-variance at the therapist level. We then carried out a factorial experiment where these factors were manipulated, and in addition, we also manipulated the amount of imbalance in the number of subjects per therapist, compared full versus partial nesting and the impact of missing data. In total, there were 1,584 different conditions evaluated.

6.2.2 Results

The empirical simulation results validated the analytical results, and showed that even when 5% of the variance in slopes is at the therapist level, the type I errors can be substantially inflated. The simulations also revealed that unbalanced allocation of patients to the therapists can have a large impact of the type I errors, when the therapist level is ignored.

6.2.3 Conclusions

Potential therapist effects can have a substantial impact on the type I errors and yield highly unreliable results. When analyzing data from longitudinal studies, investigators should account for the possibility that therapists might have different overall slopes over time. In an LMM, this can be accounted for by including a random slope at the therapist level.

6.3 Study II: "Modeling Longitudinal Gambling Data: Challenges and Opportunities"

Gambling expenditure is a common outcome in gambling research, and it is recommended as a treatment outcome in the Banff consensus statement (Walker et al., 2006). However, analyzing gambling expenditure poses many challenges. Not only is the outcome highly skewed with some participants losing a small amount of money and some very large amounts—if the treatment is successful there will also be a large proportion of reports that are zero.

6.3.1 Methods

We used data from a recent RCT comparing two behavioral interventions aimed at problem gamblers (Nilsson, Magnusson, Carlbring, Andersson, & Hellner Gumpert, 2016), to highlight the issues and show that our proposed model can be a more attractive option. We also reviewed 69 published articles to understand better how authors tend to deal with the problem. The review showed that the problem is well recognized and that researchers try to deal with the problem mostly by log transforming the outcome (most likely a $\log(y + 1)$ transformation when the outcome include zeros), or continue with a standard analysis based on a normally distributed residuals. Furthermore, we compared the performance of the proposed two-part model to the typical methods used by investigators: a linear mixed-effects model with or without a $\log(y + 1)$ transformation. The performance of these models was compared using different Monte Carlo simulation scenarios. The choice of an appropriate estimand for treatment effects was also discussed, where we argue that gambling researchers should primarily be concerned with the overall reduction in gambling losses.

6.3.2 Results

In general, the classical LMM with our without a $\log(y + 1)$ transformation were both biased and substantially less efficient (i.e., had less power) compared to the two-part

model. The $\log(y + 1)$ transformation performed worse when the proportion of zeros differed between treatment arms—which badly biased the results in some scenarios and even obscured a large overall reduction in gambling losses.

6.3.3 Conclusions

The marginalized two-part model is an attractive option to model gambling losses as a treatment outcome. Under the assumptions that gambling losses are semicontinuous, and that the conditionally positive losses follow a gamma distribution the proposed two-part model is superior to or equal to the other models as measured by either the CI's coverage probabilities, power, bias, and root-mean-square error (RMSE).

6.4 Study III: "Level of Agreement Between Problem Gamblers' and Collaterals' Reports"

In Study III, we investigated the level of agreement between problem gamblers and their CSOs. We also demonstrated the utility of the two-part model when calculating ICCs as compared to the Gaussian LMM.

6.4.1 Methods

The sample consisted of problem gamblers and their CSOs participating in a trial comparing individual CBT versus behavioral couples therapy (Nilsson et al., 2016). A total of 133 dyads were included, and we used their baseline reports of gambling losses using the timeline followback covering the last 30-days. We used a two-part model with a dyad-level random intercept and compared both a lognormal and gamma response distribution. The level of agreement was estimated using the intraclass correlation coefficient (ICC). We also compared whether the level of agreement differed as a function of the type of CSO (parent, partner, or other).

6.4.2 Results

Overall there was a fair-level of agreement, ICC = .57, 95% CI [.48, .64]. There were some evidence that partner CSOs had a higher level of agreement compared to parent CSOs, ICC_{diff} = .20, 95% CI [.03, .39].

6.4.3 Conclusions

In this study, we show two things: First, in this type of population, CSOs and problem gamblers are fairly in agreement regarding the amount of money lost. Second, ICCs

calculated using the Gaussian LMM are highly unreliable, and the nature of gambling losses make the normal assumptions highly unlikely to hold. A small simulation study both validated the two-part model, and further showed that the Gaussian model resulted in biased ICC estimates under assumptions relevant to our study. Thus, even if the ICCs are more complicated to calculate using the two-part GLMM, it can be worth the trouble since its estimates are more precise, less biased, and more informative.

6.5 Study IV: "Internet-delivered Cognitive-behavioral Therapy for Concerned Significant Others of People with Problem gambling"

In Study IV, we investigated the efficacy of an Internet-delivered CBT program for CSOs of treatment refusing problem gamblers. This study use the findings from Study I, II, and III: 1) we adjusted for therapist effects, 2) used the CSOs reports of their loved one's gambling losses 3) and used the two-part model to analyze the longitudinal reports of gambling losses.

6.5.1 Methods

In total, 100 CSOs of treatment-refusing problem gamblers were randomized to either ten weeks of ICBT or a waitlist control. The primary analysis assumed a MAR missing data mechanism; however, we performed sensitivity analyses using both pattern-mixture methods and multilevel multiple imputation. We tried to estimate the causal effect of adhering to the intervention by using data on the number of completed worksheets and the time spent on the online treatment modules.

This trial was prospectively registered with clinicaltrials.gov (NCT02250586), and a study protocol with a more detailed description of the trials is published open access (Magnusson, Nilsson, Hellner Gumpert, Andersson, & Carlbring, 2015).

For transparency and for better pooling of data, we also published the raw data, including all measured outcomes together with the R scripts used to analyze the trial. The data and scripts can be downloaded from https://osf.io/awtg7.

6.5.2 Results

At posttest the intervention group reported an improvement on the CSO's emotional consequences (d = -0.90, 95% CI [-1.47, -0.33]), relationship satisfaction (d = 0.41, 95% [0.05, 0.76]), anxiety (d = -0.45, 95% [-0.81, -0.09]), depression (d = -0.49, 95% [-0.82,

-0.16]). Any effects on the CSO's reports of gambling losses and treatment-seeking were inconclusive.

6.5.3 Conclusions

Problem gamblers are hard to influence via their CSO proxies; however, the intervention had a clinically meaningful effect of the CSO's coping as measured by their emotional consequences, anxiety, depression, and relationship satisfaction. It also seems like that there was a dose-response effect, where participants that engaged more with the intervention benefited more.

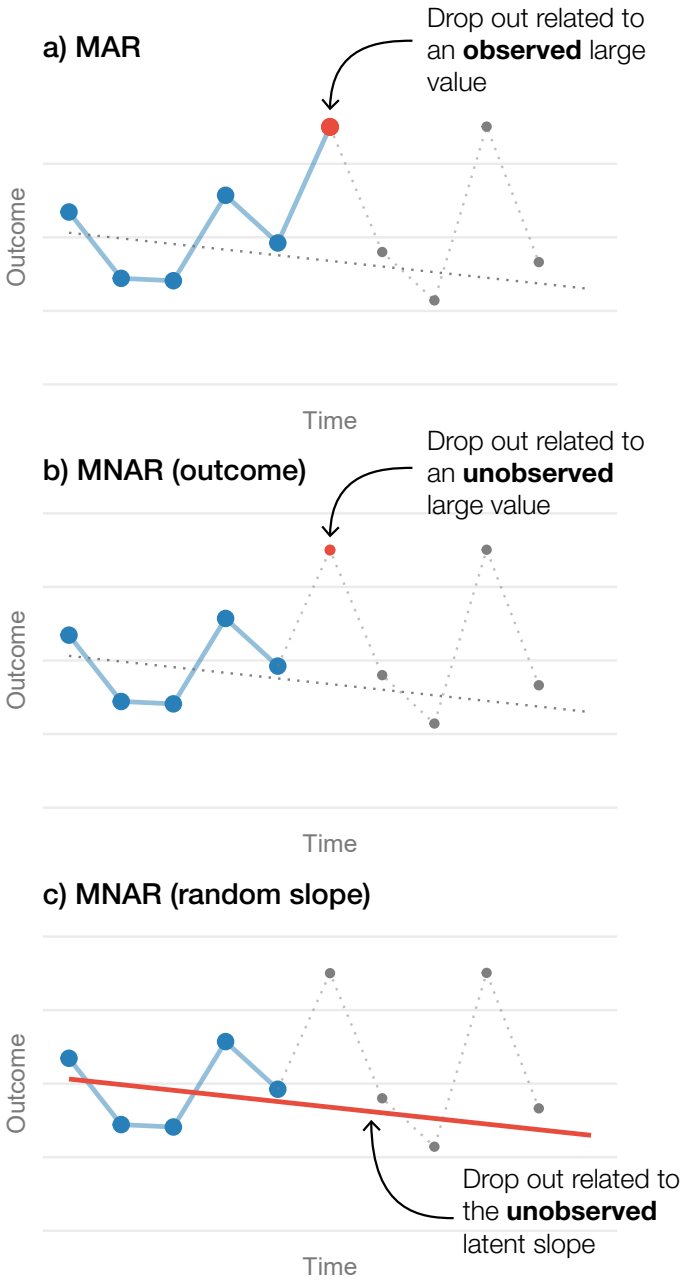


Figure 6.6: Three different drop out mechanisms in longitudinal data from one patient. a) Illustrates a MAR mechanism where the patient's likelihood of dropping out is related to an observed large value. b) Shows an outcome-related MNAR mechanism, where dropout is related to a large unobserved value. c) Shows a random-slope MNAR mechanism where the likelihood of dropping out is related to the patient's unobserved slope.

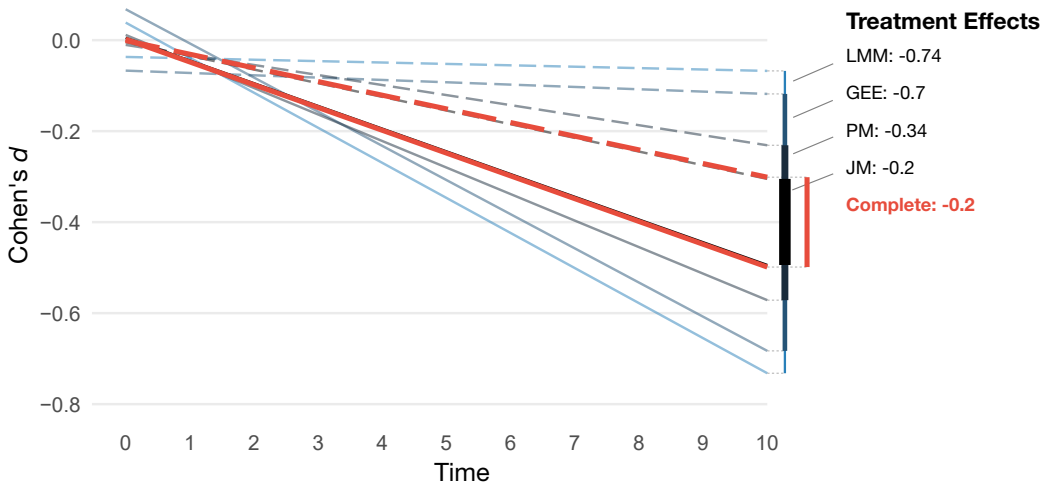


Figure 6.7: Mean of the estimated treatment effect from the MNAR missing data simulations for the different models. The dashed lines represents the control group's estimated average slope and the solid lines the treatment group's average slope.

Chapter 7

Discussion

> *"Null hypothesis testing of correlational predictions from weak substantive theories in soft psychology is subject to the influence of ten obfuscating factors whose effects are usually (1) sizeable, (2) opposed, (3) variable, and (4) unknown. The net epistemic effect of these ten obfuscating influences is that the usual research literature review is well nigh uninterpretable. Major changes in graduate education, conduct of research, and editorial policy are proposed."*
>
> – Paul E. Meehl (1990)

This discussion will start with issues more specifically related to gambling disorder interventions, and then move on to the broader topics related to psychotherapy research in general.

7.1 Helping CSOs and Problem Gamblers

Being a CSO of a problem gambler is not an easy situation, and evaluating interventions aimed towards CSOs is important. However, evaluating such interventions was challenging, and there are several aspects that deserve further scrutiny.

7.1.1 Choice of Outcome

The primary outcome in the trial was the *Inventory of Consequences Scale for the Gambler and CSO* (ICS). It consists of three sub-scales intended to measure: 1) the CSO's emotional consequences, 2) the CSO's behavioral consequences, and 3) consequences for the gambler. One could rightly ask if this is the most relevant primary outcome and that, in theory, treatment-entry is more important. However, there are several issues

with treatment-entry as an outcome as well. First, there is a lack of treatment options which might lead to an underestimation of the effect of the intervention. Second, just entering treatment does not mean that the gambling problem has gone away, and even for patients that finish the treatment, their situation might not improve much. Thus, using treatment-entry as an outcome could overestimate the actual benefits of the intervention. Third, the CSOs might benefit even without the gambler entering treatment; for instance, CSOs might be successful in influencing the problem gambler to quit or at least reduce the negative consequences caused by gambling, without the gambler having to enter treatment.

Using a scale such as the ICS to try to capture different dimensions of the gambling harms is not a bad idea. The problem with the ICS scale, however, is that it has unclear validity. We do not know if it captures a change in the consequences caused by problem gambling, and we do not know how to interpret the scores. Moreover, all items are weighted the same, but it seems highly unlikely that all consequences are equal and that we should put more weight on others.

It is also possible that the meaning of some of the items varies for different subgroups (i.e., lack of measurement invariance). For instance, the items' meaning might differ for parents and partners, or for CSOs that live in the same household as the gambler versus those who live apart. A similar concern is if the intervention actually impacts how the CSOs perceives the questions. Part of the intervention is psychoeducation about gambling problems. Hence, it is possible that the relationship between the construct and the measure change over time due to the CSOs learning more about gambling problems. This raises the question: does a 10-point difference mean the same at baseline and at posttest for the treated versus the control participants?

7.1.2 Choice of Comparator

As I discuss in the background, the waitlist control has received much criticism. However, considering the lack of previous research, the use of a waitlist control in Study IV can be defended in this trial. Moreover, the person with the problem gambling is potentially blinded as they need not know about the CSO's participation in the trial. As we discuss in the paper, this could lead to a difference in the outcomes related to the CSO and the gambler. However, it is still true that the comparator does not control for non-specific effects. The larger effect on the CSO's emotional well-being might be entirely caused by the telephone contact and the CSO's relation to the counselor. For instance, the measures of adherence could just be a proxy for the amount of time talking or messaging with the counselor.

7.1.3 Feasibility

By looking at the usage data of the online program, we can conclude that the uptake varies substantially. The written feedback indicates that there is a group of CSOs who find this type of intervention meaningful and who would recommend it. In many ways, low adherence is understandable. The CSO's are not the one engaging in the problematic behavior. Working through a program over ten weeks (partly by themselves) requires a lot of motivation, probably more than usual compared to, for instance, ICBT for social phobia or depression. The CSO has to deal with the gambler owing them money and not paying for their shared expenses, in addition, it might be hard to even notice a change in the gambling behavior if the gambler tries to conceal it. If they also work full-time and have young children, it is understandable why it might be hard to find the time and motivation to work on an online program for "someone else's" problems.

Still, some CSOs did adhere to the program, and the dose-response analysis indicated a beneficial effect. More research is needed to investigate the causal effect of sticking with the program. But it is possible that the treatment effect is clinically relevant for the participants who are motivated and actually adhere to the program. Our clinical impression during the study was that the group of participants is very heterogeneous, some were parents to gamblers living in another town, and some were spouses living with the gambler—two completely different situations.

The choice of outcome questionnaires might have negatively impacted missing data. If it feels like you are just guessing, it might not seem meaningful to return the 12-month follow-up.

7.1.4 What's the Mechanism?

The main idea behind the intervention was to first help the CSOs by focusing on their own needs via a type of behavioral activation. Then arrange for rewarding activities they could to together with the gambler. These shared activities were thought to naturally bring the gambler and CSO closer together, and introduce gambling-free activities to the gambler. After those core activities had been established, we wanted the CSO to practice how to communicate with the gambler about the gambling problem. We tried to place rather little emphasis on the CSO being able to identify the actual gambling behavior.

It seems likely that many of the CSOs spent very little time practicing the skills intended towards influencing the gambler. Personally, I think these skills would require a lot of practice, preferably by role-playing different scenarios with the therapist. It is not easy to change the dynamics in a dyad when emotions run hot. However, anecdotally, there were some CSOs who did practice these skills thoroughly and reported that they successfully

improved how they communicate about gambling.

Lastly, a difficult question is: were the CSO's emotional problems caused by the problem gambling? Most research on this topic is correlational. This question is related to the potential measurement issues discussed earlier, one can rightfully ask if the changes in the CSO's emotional consequences (ICS) and in the depression (PHQ-9) symptoms and anxiety (GAD-7) reflects a reduction in "gambling harms" and increased cooping—or is the mechanisms similar to other ICBT interventions aimed at reducing depression and anxiety? It is possible that we would have seen similar results on these outcomes if the CSOs had telephone contact with a counselor that validated them, provided some psychoeducation about gambling, and then motivated them to focus less on the gambling and instead do "rewarding and fun activities".

Clearly, these causal mechanisms will be heterogeneous. Some CSOs showed very little emotional distress and did not suffer financial consequences—others were more directly affected. It is possible that adherence and maybe outcomes could be improved if the intervention is more tailored toward the CSOs specific situation. The original CRAFT approach is more of a "smörgåsbord" with a more flexible treatment-planning than our online intervention.

7.1.5 Agreement Between the CSO and Gambler

Study III provided some evidence that it is possible for CSOs and gamblers to show a decent level of agreement in regards to gambling losses. It is likely that CSOs who cohabit and have shared expenses have a better grasp on the losses—at least large relative changes in losses. If the gambler reduces or quits their gambling, the CSO could notice that the gambler can pay expenses and needs to borrow less money. Still, we know that problem gamblers have managed to hide large losses for a long time from their loved ones.

The obvious difference between the sample of gamblers and CSOs in Study III and Study IV affect the transportability of the findings. The gamblers in Study III were willing to begin treatment together with their CSOs, whereas the problem gamblers in Study IV were treatment-refusing, making it plausible that they were actively trying to hide their losses, i.e., the agreement would probably be even lower in such a sample. Currently, there has been no research on the validity of collateral reports in such a sample.

Moreover, we only investigated agreement using the baseline reports, since we wanted to avoid that the reports were influenced by the treatment. From a treatment point-of-view, it would be interesting to see if agreement increases over time for dyads receiving behavioral couples therapy versus CBT. However, the large number of missing

observations will make such an analysis extremely hard to interpret.

7.1.6 Clinical Implications

The results in Study IV implies that CSOs can learn to cope better with the problem gambling. However, we still know very little about how to help CSOs influence the actual problem gambling or treatment-seeking behavior. Moreover, it is unclear if an internet intervention is the preferred mode of delivery among CSOs. However, access to other options will likely not become available nationwide in the foreseeable future. A reduced version of our 10-week program might be worth evaluating in future studies.

7.2 Modeling Gambling Losses

It is hard to image that gambling losses fulfill the typical LMM assumptions, i.e., normally distributed residuals with a constant variance, an additive effect of treatment, and multivariate normal random effects. The longitudinal two-part model proposed in Study II is *a priori* much more plausible. The model can easily capture the fact that some will completely stop gambling, and some will continue to gamble heavily. However, many challenges remain when it comes to analyzing gambling losses as a treatment outcome. Some of the issues are:

- Clinical significance: It is hard to define what a clinically significant improvement would be. For some, a reduction in losses of 100 SEK per day would be clinically relevant; for others, it would make no difference. Although we could use the relative units an say that a 20% reduction is clinically relevant, it is not clear that this is any more clinically meaningful.
- Recovery: A related issue is to try to define "recovery" from problem gambling. Although, I consider the focus on "recovery" as a treatment outcome, an unnecessary dichotomization of a continuous outcome (gambling losses). Undoubtedly, a problem gambler will still have residual problems even if they completely stop gambling, such as financial problems, gambling urges, and relationship problems. However, it seems natural that the primary aim of a gambling intervention should be to reduce the problematic gambling. Thus, when evaluating the efficacy of a treatment, what matters is if the treatment increases the likelihood of staying abstinent, or lead to a reduction in the gambling losses. In clinical practice, other significant comorbidities that patients might suffer from can likely be targeted by combining treatments.
- Dropout: Some participants will drop out because they feel they are doing fine and do not need any more treatment. For these participants, dropout might depend

on the observed values (the MAR assumption). For others, they might drop out because they relapse, or because they feel like the treatment is not helping or they might not like their therapist and feel like they do not understand their gambling problem. Whatever the reason is, the dropout process is likely to be a mix of MAR and MNAR related mechanisms. Dropout will likely continue to be a major challenge in gambling studies. What we can improve is to think harder about appropriate sensitivity analyses. We can also include questions if the participant think they will drop out before the next visit—information that could make the MAR assumption more plausible.

- Long term effects: We know that study participants tend to stop gambling once they enter treatment. Naturally, this change can not be caused by the contents of the treatment and should be seen as an effect of the participant's resolve to change their behavior. Deciding to enter treatment might be enough to keep them motivated to abstain from gambling during the treatment period. Thus, the real treatment effect would be the impact on the long-term outcomes, say, 5 to 10 years later. Clearly, the effect of randomization will be long broken by that time, and dropout will be an even greater issue. Still, preventing relapse, in the long run, might be where differences between treatments would show up.

In addition to the issues just listed, it is not entirely clear how reliable retrospective reports of gambling losses are. However, more and more gambling is tracked, and gamblers can see how great their losses are. Thus, EMA or diary-type reports can be even further improved by the aid of the tracked gambling data. The gambling field might be one of the few clinical fields were the possibility exists of using behavioral tracking to receive ecologically valid data on the problem behavior. For some type of studies, it is even possible to collaborate with the gambling industry and have access to transactional-level data. However, such collaborations are not uncontroversial; the gambling industry does not always share the agenda of the researchers.

Lastly, one of the major challenges in Study II is how to disseminate the proposed model. I doubt it can be fit using SPSS, and even in R or SAS, it is not straight-forward and most researchers might need to consult an expert. This is a major drawback for the dissemination of a method.

7.3 Therapist Effects

Having discussed specific issues related to gambling studies, I will now focus on the broader issue of therapist effects in treatment studies.

7.3.1 Therapists, Do They Matter?

Little is known about how strong the therapist effect is, mostly because researchers continue to use statistical models that blatantly ignore the existence of therapists. Naturally, it is likely that the variance at the therapist level is much smaller compared to the variance between patients. However, as we showed in Study I, the consequences of ignoring this variance can have large consequences for the statistical tests.

There are many reasons why therapists might differ in their overall success. Some might just have more skill on the relevant variables—whether that is their ability to create expectations and alliance, or their adhererence to the specific ingredients of the treatment. Some therapists might just underperform during the study; things could happen in their private lives that affect their work, or there might be a problem at the organizational level creating unrealistic working conditions.

It is important to note that therapist effects do not make the treatment effect invalid. If the therapists in each arm are comparable, then the treatment effect will not be biased; it will simply shift the whole distribution of therapists giving the treatment. Thus, the difference between therapists at the 95th percentile in either treatment group is the same as for therapists at the median in either treatment group. However, if the variances are heterogeneous, i.e., different in each arm, then the therapists will be a source of treatment effect heterogeneity. This is also the case for a partially nested design, where, indeed, the treatment effect is larger for a therapist at the 95th percentile compared to the expected outcome in the control group. Figure 7.1 illustrate these three different scenarios.

An important question to ask is: Why do these distributions differ? In a nested design, the therapists and treatments are confounded. The therapists giving treatment A could be successful because they are more competent, which would inflate the estimated treatment effect. This could be caused by the fact that a treatment that is more popular and seen as more evidence-based might attract more ambitious and skilled therapists as they know that this will improve their work opportunities. The relationship could also be the reverse, i.e., there is no real difference between the therapists' giving either treatment. In this case, the overall difference mostly reflects the effect of the specific treatment. Moreover, it is possible that there is no overall therapist effect, in the sense that therapists would perform well using any treatment due to, for instance, their mastery of common factors. In a nested design, it is not possible to partition out if some of the variance is caused by a therapist $\times$ treatment interaction. It is possible that all of the variance stems from an interaction between therapist characteristics and the treatment—and that a "super therapist" only outperforms their peers when delivering their preferred treatment. Clearly, all of this has a bearing on whether therapists or

treatments are more important.

One could also look at Figure 7.1 and note that the overlap is substantial. There will be therapists from the less beneficial treatment B that outperforms therapists from treatment A. Based on such reasoning, some would claim that therapists are more important than the treatment provided. However, it makes little sense to compare therapists at different percentiles—and as noted earlier we do not know if it is the therapists causing the treatment effects or if it is the treatment effects causing the difference between therapists. However, if the distribution of the expected outcomes is valid (i.e., it approximates the empirical distribution of the real therapists), then all of these therapists will treat patients. Thus at the population level, our best guess would be to pick treatment A, either because the treatment is better or because the therapists delivering the treatment tend to be better. Clearly, if I was a patient and I had access to this information, my best option would be to pick one of the top therapists giving treatment A. Unfortunately, I do not have access to this information, and no one else has, so again, my best option would be to pick a therapist from treatment A.

7.3.2 Fixed Versus Random Effects Again

The discussion regarding if therapists should be viewed as a fixed or random effect is somewhat strange. The argument for random effects (varying effects) is about generalizability. The confidence intervals will be wider for the treatment effect, as we assume therapists are sampled from a distribution. This makes sense if we are thinking about effectiveness. However, if we are primarily concerned with estimating the efficacy of a treatment, then conditioning on the therapists in the study makes sense. A problem is that psychotherapy trials do not really have different phases, in the same way that pharmacological trials do. It would make sense to use a fixed effects approach in the early phases, where power is a major concern and where a type II error could lead to abandoning research on a treatment that is indeed efficacious. Before dissemination to the general public, larger studies could be performed where therapists are a random factor.

A third option is to perform a fixed effects analysis but adjust the standard errors, a method more common in econometrics. There is no gain in power in doing this, but it can be an attractive option if there is a concern about the validity of the random effects (for instance, endogeneity concerns).

7.3.3 Identifying Factors That Explain Therapist Variance

If we can identify variables that explain some of the variance between therapists, such as competence or experience, then measuring and including these variables in

our models should increase the model's precision by reducing the between-therapist variance. Considering that most studies completely ignore therapist effects, this means that very few good markers exist. Such problems can easily be avoided by publishing de-identified data. Instead, thousands of studies have been performed that contained useful information that could have been used to design better studies—unfortunately, one can guess that most of the data sets are lost forever.

7.4 Psychotherapy Research—Looking Forward

After having spent most of this thesis looking at the bad parts of psychotherapy research, I will end this discussion by focusing on how things can be improved in the future.

7.4.1 Methods Issues and Dissemination

A wide-spread problem in research is that calls of concern from methodologists continue to be ignored. For example, the consequences of ignoring therapist effects have been known for a long time, with little impact on the way investigators analyze and report their trials. It is clear that just writing method papers is not enough to improve our field. Most researchers in clinical psychology both lack strong quantitative training, programming knowledge or have the prerequisite mathematical knowledge that they would need to apply or evaluate many method papers. With the challenges often faced in clinical psychology, it should not be surprising that clinical researchers struggle with using modern computational tools. However, I do not think that the solution is to focus solely on the individual researcher's quantitative training. It is not reasonable to expect clinical researchers to become quantitative and clinical experts simultaneously. The issues we are facing need to be dealt with by reforming how we organize, reward, and fund research. With that being said, I do think that the overall quantitative knowledge needs to be improved. Not so that everyone should be able to code their own custom analyses, but rather so that clinical investigators can ask better questions and better evaluate the limitations and inferences in their trials.

The problem of disseminating quantitative methods shares many similarities with the research-to-practice gap in healthcare, and several lessons from implementation sciences can most likely apply to the research-to-practice gap in quantitative methods (King, Pullmann, Lyon, Dorsey, & Lewis, 2019).

A specific issue that pertains to this thesis is how to disseminate method papers best. Even the applied quantitative papers tend to be too technical for the typical psychotherapy researcher. In many ways, quantitative scholars face a user interface (UI) and user experience (UX) design problem. If a quantitative researcher wants to publish

a new method (a new "product") that researchers should use, then just publishing a method paper, tend to lead to really sub-par UI and UX. Very few researchers will use a method unless it is packaged in a user-friendly way, and unless the product includes the features that match their needs. In my work, I have at least tried to go beyond just publishing papers. I have tried to include more visual material and more user-friendly alternatives. For instance, is my belief that many concepts can be explained better using interactive visualizations (e.g., https://rpsychologist.com/d3/CI/), than just static pictures or using formulas. For Study I, I created my R package `powerlmm`, which can perform everything mentioned in the article (and more), and I included a web application (a Shiny app) where the core features have a more user-friendly graphical interface.

7.4.2 Causal Inference, Learning to Let Go of Experiments

Psychotherapy investigators tend to be most familiar with one of the most basic experiments: the parallel-group randomized trial, where the only experimental manipulation is the offering of recieving treatment or not. Often we fail to recognize that many of the questions we are interested in are not experimentally manipulated, such as process measures, per-protocol effects, or dose-response relationships. Which can lead to investigators choosing analyses that are so naive that no one would have faith in the results if all assumption were explicitly stated. Going forward we should use our substantive knowledge and clinical experience together with modern causal inference methods to try to discover causal effects with more realistic models—instead of wasting our efforts on unrealistic mediation or predictive models.

We could also focus on basic science experiments to try to identify potential variables to target in our treatments. For the reasons covered in Section 2.2.5, I am personally not that optimistic that such an approach will substantially inform the practice of psychotherapy. There is a concern that more novel and hyped experimental findings will lead to trials with little hope of improving clinical outcomes, leading to an unnecessary waste of research resources (Cristea & Florian, 2019).

7.4.3 Predictive Modeling Without Buzzwords

It is understandable that psychotherapy researchers, like so many others, have a hard time resisting the hype of machine learning or "artificial intelligence" as a way of improving treatment selection, predicting who will respond to treatment, or classifying non-responders during the treatment. Indeed, labels such as "deep neural nets", "reinforcement learning", or "artificial intelligence" does sound more state-of-the art than "linear regression". Unfortunately, few problems in psychotherapy research are

solved by these models, and one can question the utility of models that gained their popularity in situations with a high signal-to-noise ratio, such as in pattern recognition problems, natural language processing, or building recommendation systems (i.e., "you might also be interested in product X"). Anyone who have seen a decent amount of psychotherapy data will know that our situation is quite the opposite.

I have seen several examples of both research funders and investigators that have bought the hype of machine learning, unfortunately, with little knowledge of how to appreciate what the good parts are—and how they apply to psychotherapy research. Although, one can rightfully argue that many of the "machine learning" methods that might work well for psychotherapy researchers are just "statistics"; such as cross-validation, appropriate performance measures, penalized regression, and so on.

The literature on clinical prediction models is much larger in medicine compared to psychotherapy research. Undoubtedly, *a lot* could be improved with regards to how prediction or classification problems are handled in psychotherapy research. The first would be to recognize what can actually be predicted (as covered in the background section). Using patients' improvement from baseline and trying to build a model that will predict treatment response is doomed to fail. What would help is to instead build better models for predicting patients' prognosis. It is not a surprise that, in medicine, it is with image recognition that machine learning algorithms have found their success, e.g., detecting diabetic retinopathy based on retinal photographs (Beam & Kohane, 2018; Gulshan et al., 2016).

7.4.4 "We Need Less Research, Better Research, and Research Done for the Right Reasons"

Many of the issues I discuss would require changes to the way research is undertaken. Hopefully, the days of psychotherapy trials being basically a solo venture are soon gone; were one or two persons design the trial, carry out the treatment, analyze the results, and publish the results with very little transparency or oversight. Clearly, this is problematic even when investigators are conscientious and report everything transparently—no one is an expert in all of the relevant skills.[1]

Instead of many small trials of this type, run by a single academic investigator, what we need is large collaborations that include multiple institutions and with enough resources to create research teams and support functions to run large high-quality trials. This would allow teams to focus on: a) measurement problems *before* starting a trial, b) evaluate interventions in several steps, starting with single-case experimental

[1] No, not even Paul Meehl.

designs and where patients and clinicians can provide feedback on the contents of interventions and if relevant outcomes are captured, c) include outcome measures that are actually validated specifically for the problem being investigated, d) letting clinicians, methodologists, psychometricians, and statisticians work together from the start both to formulate relevant questions and methods to answer them. Cuijpers et al. (2018b) reasoned similarly:

> "It is as if we have been in the pilot phase of research for five decades without being able to dig deeper. If we want to take a step forward, we need to conduct research that goes beyond examining, on the one hand, simple correlational associations between specific and common factors and, on the other, outcomes … There are no easy solutions, and such research will require considerable resources. However, we have invested resources in this research for five decades, and if we could put only part of these resources toward making a coordinated effort to examine mechanisms of change, it would certainly become feasible" (p. 18)

Moreover, in order to realize this scientific utopia, researchers would need to publish *less* in order to have the time to produce reliable research that is carefully thought through and carried out. Obviously, this would require changing the incentive structures in science and how researchers are promoted and recruited—a pretty daunting task.

7.4.5 Embrace Open Science

In order to create a cumulative and transparent research field data and scripts need to be open, and psychotherapy researchers need to embrace open science practices. It is a clear research waste that so much data on psychotherapy outcomes and processes are unavailable, and perhaps, permanently lost. Open data and scripts also enable science to be self-correcting; currently, it is close to impossible to check claims made in published trials.

Science is not open if only those with the appropriate software licenses can run our code—open science is best done using open software. Naturally, treatment manuals and worksheets should be published using a permissive license so that others are free to use and adapt the materials. These are simple steps that would improve the trustworthiness of psychotherapy research substantively. With most of these items included in the most recent CONSORT statement for social and psychological interventions (CONSORT-SPI; Grant et al., 2018), maybe the situation will improve. For instance, item 12a states:

> "To facilitate full reproducibility, authors should report software used to run analyses and provide the exact statistical code" (p. 11)

and item 17a:

> "As part of the growing open-science movement, triallists are increasingly expected to maintain their datasets, linked via trial registrations and posted in trusted online repositories … to facilitate reproducibility of reported analyses and future secondary data analyses" (p. 13)

Hopefully, psychotherapy researchers or funders recognize the importance of these items.

7.4.6 Is it Time to Regulate Psychotherapy Research?

It is quite strange that psychotherapies (or "psychological treatments") can be offered by the Swedish (and other) healthcare systems without any specific regulation of the psychotherapy research which is used to guide what treatments are offered.

It should be evident that the reporting of psychotherapy trials in academic journals is a sub-optimal way of ensuring quality. When "psychological treatments" are included in the healthcare system and provided by licensed psychologists, the evidence needs to be evaluated independently by a responsible governing body. We cannot rely on simply the published literature and base healthcare recommendations on systematic reviews of this literature. A better approach would be to adopt the system for introducing new medicines on the market where there is a governing body and specific regulation, such as the *European Medicine Agency* and the Swedish *Läkemedelsverket*. In order to reclaim trust in the evaluation of psychotherapy interventions, trial data need to be independently verified, and regulatory documents should detail how the research should be conducted and how *good clinical practice* is ensured. After a psychotherapy is shown to be effective, its implementation needs to be monitored, and guidelines are needed for how to train therapists and what competencies are required.

7.5 Concluding Remarks

Undoubtedly, the challenges faced by psychotherapy researchers are monumental. It is easy to sound negative when focusing mostly on research issues. Hopefully, improvements to psychotherapy research will follow, and that these improvements will improve clinical practice and reduce the mental health burden in general.

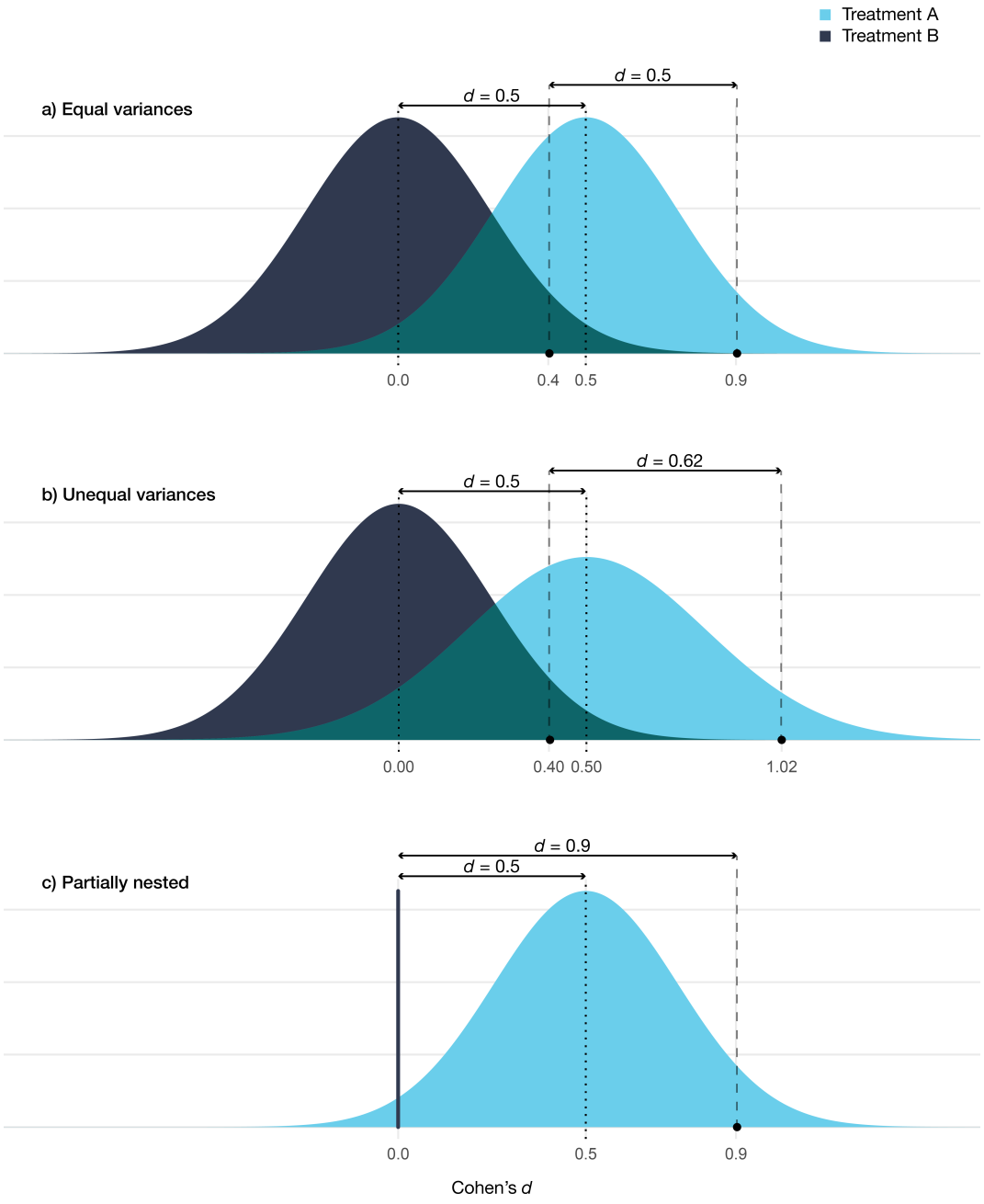


Figure 7.1: Distribution of therapists' expected outcome for two different treatments. In a) the variance is equal in both treatments (at posttest ICC = 0.05), in b) there is a larger variance in treatment A (at posttest ICC = 0.05 vs 0.08), in c) treatment B is a no-treatment control condition (at posttest ICC = 0.05 in treatment A). Comparison between treatments are shown at the median and at the 95th percentile.

Acknowledgments

A large part of this thesis focuses on quantitative methods, but one must never forget that the numbers and models represent real lives and real suffering. Numbers are meaningless if they do not correspond to real change. Therefore, I want to thank all the participants in our studies. I am especially grateful to everyone who persisted with our internet intervention even though progress might have felt slow, and our measures were too long and hard to answer. You have contributed to increasing our understanding of problem gambling even if you might not have felt like it.

Thanks to everyone involved in the clinical trials, all the staff, and all the clinicians.

Thanks to all colleagues and collaborators both at CPF and SU, there have been too many interesting people coming and going over the years. I do not dare to try and name everybody. A special thanks to Maria Garke and Viktor Månsson for helping with proofreading the "kappa" (thesis frame).

I would also like to thank all the people who have sent me random encouragement, questions, feedback, feature requests, bug reports, or who have shared my work. Doctoral studies can be a solitary activity with little feedback from the outside world. I know I have missed replying to many of you, but I honestly feel that you have made these years much more joyful.

A big thanks to all the people all around the world, sacrificing their nights and weekends to create open-source software that improves science. You do not get enough credit for your work.

Lastly, thanks to my supervisors, Per Carlbring, Clara Hellner, and Gerhard Andersson. A special thanks to Per, my main supervisor, for always being supportive and giving me the freedom to explore new ideas and follow my interests.

References

Abbott, M., Romild, U., & Volberg, R. (2018). The prevalence, incidence, and gender and age-specific incidence of problem gambling: Results of the Swedish longitudinal gambling study (Swelogs). *Addiction*, *113*(4), 699–707. https://doi.org/10.1111/add.14083

Abbott, M. W., Romild, U., & Volberg, R. A. (2014). Gambling and Problem Gambling in Sweden: Changes Between 1998 and 2009. *Journal of Gambling Studies*, *30*(4), 985–999. https://doi.org/10.1007/s10899-013-9396-3

Abdel-Ghany, M., & Sharpe, D. L. (2001). Lottery Expenditures in Canada: Regional Analysis of Probability of Purchase, Amount of Purchase, and Incidence. *Family and Consumer Sciences Research Journal*, *30*(1), 64–78. https://doi.org/10.1177/1077727X01301003

Altman, D. G. (1994). The scandal of poor medical research. *BMJ*, *308*(6924), 283–284. https://doi.org/10.1136/bmj.308.6924.283

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485. https://doi.org/10.1136/bmj.311.7003.485

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, *27*(1), 166–177. https://doi.org/10.1037/a0029508

Azar, M., Riehm, K. E., McKay, D., & Thombs, B. D. (2015). Transparency of Outcome Reporting and Trial Registration of Randomized Controlled Trials Published in the Journal of Consulting and Clinical Psychology. *PLOS ONE*, *10*(11), e0142894. https://doi.org/10.1371/journal.pone.0142894

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *American Journal of Psychiatry*, *161*(12), 2163–2177. https://doi.org/10.1176/appi.ajp.161.12.2163

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., … Watson, J. (2011). Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy*, *40*(1), 15–33. https://doi.org/10.1080/16506073.2010.520731

Bandyopadhyay, D., DeSantis, S. M., Korte, J. E., & Brady, K. T. (2011). Some Considerations for Excess Zeroes in Substance Abuse Research. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 376–382. https://doi.org/10.3109/00952990.2011.568080

Barlow, D. H. (2004). Psychological treatments. *American Psychologist*, *59*(9), 869–878. https://doi.org/10.1037/0003-066X.59.9.869

Barlow, D. H., Bullis, J. R., Comer, J. S., & Ametaj, A. A. (2013). Evidence-Based Psychological Treatments: An Update and a Way Forward. *Annual Review of Clinical Psychology*, *9*(1), 1–27. https://doi.org/10.1146/annurev-clinpsy-050212-185629

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal Stratification Approach to Broken Randomized Experiments. *Journal of the American Statistical Association*, *98*(462), 299–323. https://doi.org/10.1198/016214503000071

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Barry, D. T., Stefanovics, E. A., Desai, R. A., & Potenza, M. N. (2011). Differences in the associations between gambling problem severity and psychiatric disorders among black and white adults: Findings from the National Epidemiologic Survey on Alcohol and Related Conditions. *The American Journal on Addictions*, *20*(1), 69–77. https://doi.org/10.1111/j.1521-0391.2010.00098.x

Basu, A., & Manning, W. G. (2009). Issues for the Next Generation of Health Care Cost Analyses. *Medical Care*, *47*(7_Supplement_1), S109. https://doi.org/10.1097/MLR.0b013e31819c94a1

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*. https://doi.org/10.1001/jama.2017.18391

Bellringer, M., Pulford, J., Abbott, M., DeSouza, R., & Clarke, D. (2008). Problem Gambling - Barriers to Help-seeking Behaviours (Final Report). Retrieved from http://aut.researchgateway.ac.nz/handle/10292/2014

Bertrand, K., Dufour, M., Wright, J., & Lasnier, B. (2008). Adapted Couple Therapy (ACT) for Pathological Gamblers: A Promising Avenue. *Journal of Gambling Studies*, *24*(3), 393. https://doi.org/10.1007/s10899-008-9100-1

Beutler, L. E. (1991). Have all won and must all have prizes? Revisiting Luborsky et al.'s verdict. *Journal of Consulting and Clinical Psychology*, *59*(2), 226–232. https://doi.org/10.1037/0022-006X.59.2.226

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Bouwmeester, W., Zuithoff, N. P. A., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., … Moons, K. G. M. (2012). Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLOS Medicine*, *9*(5), e1001221. https://doi.org/10.1371/journal.pmed.1001221

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building, 201–236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Bradley, H. A., Rucklidge, J. J., & Mulder, R. T. (2017). A systematic review of trial registration and selective outcome reporting in psychotherapy randomized controlled trials. *Acta Psychiatrica Scandinavica*, *135*(1), 65–77. https://doi.org/10.1111/acps.12647

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550–558. https://doi.org/10.1037/a0018933

Burns, D. D., & Spangler, D. L. (2000). Does psychotherapy homework lead to improvements in depression in cognitive-behavioral therapy or does improvement lead to increased homework compliance? *Journal of Consulting and Clinical Psychology*, *68*(1), 46–56.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. https://doi.org/10.1002/sim.2673

Carlbring, P., Jonsson, J., Josephson, H., & Forsberg, L. (2010). Motivational Interviewing Versus Cognitive Behavioral Group Therapy in the Treatment of Problem and Pathological Gambling: A Randomized Controlled Trial. *Cognitive Behaviour Therapy*, *39*(2), 92–103. https://doi.org/10.1080/16506070903190245

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86–89. https://doi.org/10.1016/S0140-6736(09)60329-9

Chambless, D. L., & Hollon, S. D. (2012). Treatment validity for intervention studies. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (pp. 529–552). Washington, DC, US: American Psychological Association. https://doi.org/10.1037/13620-028

Clarke, D., Abbott, M., DeSouza, R., & Bellringer, M. (2007). An Overview of Help Seeking by Problem Gamblers and their Families Including Barriers to and Relevance of Services. *International Journal of Mental Health and Addiction*, *5*(4), 292–306. https://doi.org/10.1007/s11469-007-9063-y

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*(1), null. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Copello, A. G., Velleman, R. D. B., & Templeton, L. J. (2005). Family interventions in the treatment of alcohol and drug problems. *Drug and Alcohol Review*, *24*(4), 369–385. https://doi.org/10.1080/09595230500302356

Cowlishaw, S., Merkouris, S., Dowling, N., Anderson, C., Jackson, A., & Thomas, S. (2012). Psychological therapies for pathological and problem gambling. *The Cochrane Database of Systematic Reviews*, *11*, CD008937. https://doi.org/10.1002/14651858.CD008937.pub2

Coyne, J. C., & Kok, R. N. (2014). Salvaging psychotherapy research: A manifesto. *Journal of Evidence-Based Psychotherapies*, *14*(2), 104–124.

Cristea, I. A. (2018). The waiting list is an inadequate benchmark for estimating the effectiveness of psychotherapy for depression. *Epidemiology and Psychiatric Sciences*, 1–2. https://doi.org/10.1017/S2045796018000665

Cristea, I., & Florian, N. (2019). Increase value and reduce waste in research on psychological therapies. https://doi.org/10.31219/osf.io/ps7x2

Cristea, I., & Ioannidis, J. P. A. (2018). Improving Disclosure of Financial Conflicts of Interest for Research on Psychosocial Interventions. *JAMA Psychiatry*, *75*(6), 541–542. https://doi.org/10.1001/jamapsychiatry.2018.0382

Crits-christoph, P., & Mintz, J. (1991). Implications of Therapist Effects for the Design and Analysis of Comparative Studies of Psychotherapies. *Journal of Consulting and Clinical Psychology*, *59*(1), 20–26. https://doi.org/10.1037/0022-006X.59.1.20

Crits-Christoph, P., Tu, X., & Gallop, R. (2003). Therapists as fixed versus random effects-some statistical and conceptual issues: A comment on Siemer and Joormann (2003). *Psychological Methods*, *8*(4), 518–523. https://doi.org/10.1037/1082-989X.8.4.518

Crowley, F., Eakins, J., & Jordan, D. (2012). Participation, Expenditure and Regressivity

in the Irish Lottery: Evidence from Irish Household Budget Survey 2004/2005. *The Economic and Social Review*, *43*(2, Summer), 199–225–199–225. Retrieved from https://www.esr.ie/article/view/47

Cuijpers, P., & Cristea, I. (2016). How to prove that your therapy is effective, even when it is not: A guideline. *Epidemiology and Psychiatric Sciences*, *25*(05), 428–435. https://doi.org/10.1017/S2045796015000864

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*, *15*(3), 245–258. https://doi.org/10.1002/wps.20346

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2018a). Is psychotherapy effective? Pretending everything is fine will not help the field forward. *Epidemiology and Psychiatric Sciences*, 1–2. https://doi.org/10.1017/S204579601800080X

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: A meta-analysis. *Journal of Affective Disorders*, *159*, 118–126. https://doi.org/10.1016/j.jad.2014.02.026

Cuijpers, P., Reijnders, M., & Huibers, M. J. H. (2018b). The Role of Common Factors in Psychotherapy Outcomes. *Annual Review of Clinical Psychology*. https://doi.org/10.1146/annurev-clinpsy-050718-095424

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010a). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry*, *196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *31*(5), 374–378. https://doi.org/10.1002/da.22249

Cuijpers, P., van Straten, A., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010b). The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size. *Psychological Medicine*, *40*(2), 211–223. https://doi.org/10.1017/S0033291709006114

Cunningham, J. A. (2005). Little Use of Treatment Among Problem Gamblers. *Psychiatric Services*, *56*(8), 1024–a. https://doi.org/10.1176/appi.ps.56.8.1024-a

Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and reproducibility through registration: The status of intervention trials published

in clinical psychology journals. *Journal of Consulting and Clinical Psychology*, *84*(9), 753–767. https://doi.org/10.1037/ccp0000115

Deb, P., & Trivedi, P. K. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of Health Economics*, *21*(4), 601–625. https://doi.org/10.1016/S0167-6296(02)00008-5

de Jong, K., Moerbeek, M., & van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: An example with therapist effects. *Psychotherapy Research*, *20*(3), 273–284. https://doi.org/10.1080/10503300903376320

DeSantis, S. M., Bandyopadhyay, D., Baker, N. L., Randall, P. K., Anton, R. F., & Prisciandaro, J. J. (2013). Modeling longitudinal drinking data in clinical trials: An application to the COMBINE study. *Drug and Alcohol Dependence*, *132*(1-2), 244–250. https://doi.org/10.1016/j.drugalcdep.2013.02.013

Devereaux, P., Bhandari, M., Clarke, M., Montori, V. M., Cook, D. J., Yusuf, S., … others. (2005). Need for expertise based randomised controlled trials. *Bmj*, *330*(7482), 88.

Dickson-Swift, V. a, James, E. L., & Kippen, S. (2005). The experience of living with a problem gambler: Spouses and partners speak out. *Journal of Gambling Issues*, *13*(13), 1–22. https://doi.org/10.4309/jgi.2005.13.6

Dieleman, J. L., & Templin, T. (2014). Random-effects, fixed-effects and the within-between specification for clustered data in observational health studies: A simulation study. *PLOS ONE*, *9*(10), e110257. https://doi.org/10.1371/journal.pone.0110257

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., & others. (2002). *Analysis of longitudinal data*. Oxford University Press.

Diskin, K. M., & Hodgins, D. C. (2009). A randomized controlled trial of a single session motivational intervention for concerned gamblers. *Behaviour Research and Therapy*, *47*(5), 382–388. https://doi.org/10.1016/j.brat.2009.01.018

Driessen, E., & Hollon, S. D. (2010). Cognitive Behavioral Therapy for Mood Disorders: Efficacy, Moderators and Mediators. *Psychiatric Clinics*, *33*(3), 537–555. https://doi.org/10.1016/j.psc.2010.04.005

Driessen, E., Hollon, S. D., Bockting, C. L. H., Cuijpers, P., & Turner, E. H. (2015). Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-Funded Trials. *PLOS ONE*, *10*(9), e0137864. https://doi.org/10.1371/journal.pone.0137864

Driessen, E., Van, H. L., Don, F. J., Peen, J., Kool, S., Westra, D., … Dekker, J. J. (2013).

The Efficacy of Cognitive-Behavioral Therapy and Psychodynamic Therapy in the Outpatient Treatment of Major Depression: A Randomized Clinical Trial. *American Journal of Psychiatry*, *170*(9), 1041–1050. https://doi.org/10.1176/appi.ajp.2013.12070899

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business & Economic Statistics*, *1*(2), 115–126. https://doi.org/10.1080/07350015.1983.10509330

Dunn, G., & Bentall, R. (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, *26*(26), 4719–4745. https://doi.org/10.1002/sim.2891

Dunn, G., Maracy, M., & Tomenson, B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Statistical Methods in Medical Research*, *14*(4), 369–395. https://doi.org/10.1191/0962280205sm403oa

Echeburúa, E., Fernández-Montalvo, J., & Báez, C. (2001). Predictors of therapeutic failure in slot-machine pathological gamblers following behavioural treatment. *Behavioural and Cognitive Psychotherapy*, *29*(3), 379–383. https://doi.org/10.1017/S1352465801003113

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. *Clinical Psychology: Science and Practice*, *6*(1), 10–32.

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH treatment of depression collaborative research program: Background and research plan. *Archives of General Psychiatry*, *42*(3), 305–316.

Emsley, R., Dunn, G., & White, I. R. (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research*, *19*(3), 237–270. https://doi.org/10.1177/0962280209105014

Eysenck, H. J. (1952). The Effects of Psychotherapy: An Evaluation. *Journal of Consulting Psychology*, *16*(5), 319–324.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*(5), 517–517. https://doi.org/10.1037/0003-066X.33.5.517.a

Falkenström, F., Markowitz, J. C., Jonker, H., Philips, B., & Holmqvist, R. (2013). Can psychotherapists function as their own controls? Meta-analysis of the "crossed therapist" design in comparative psychotherapy trials. *The Journal of Clinical Psychiatry*, *74*(5), 482.

Farrell, L., & Walker, I. (1999). The welfare effects of lotto: Evidence from the UK. *Journal of Public Economics*, *72*(1), 99–120. https://doi.org/10.1016/S0047-2727(98)00089-9

Fernandez, A. C., Begley, E. A., & Marlatt, G. A. (2006). Family and peer interventions for adults: Past approaches and future directions. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *20*(2), 207–213. https://doi.org/10.1037/0893-164X.20.2.207

Fine, J. P., & Pencina, M. J. (2015). On the Quantitative Assessment of Predictive Biomarkers. *JNCI: Journal of the National Cancer Institute*, *107*(8). https://doi.org/10.1093/jnci/djv187

Fisher, A. J. (2015). Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology*, *83*(4), 825–836. https://doi.org/10.1037/ccp0000026

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Flint, J., Cuijpers, P., Horder, J., Koole, S. L., & Munafò, M. R. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological Medicine*, *45*(2), 439–446. https://doi.org/10.1017/S0033291714001421

Folkhälsomyndigheten. (2019). *Resultat från Swelogs 2018*. Statens folkhälsoinstitut. Östersund.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Furukawa, T. A., Noma, H., Caldwell, D. M., Honyashiki, M., Shinohara, K., Imai, H., … Churchill, R. (2014). Waiting list may be a nocebo condition in psychotherapy trials: A contribution from network meta-analysis. *Acta Psychiatrica Scandinavica*, *130*(3), 181–192. https://doi.org/10.1111/acps.12275

Gaffan, E. A., Tsaousis, I., & Kemp-Wheeler, S. M. (1995). Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *63*(6), 966–980.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gelman, A. (2013). *The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gold, S. M., Enck, P., Hasselmann, H., Friede, T., Hegerl, U., Mohr, D. C., & Otte, C. (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: A decision framework. *The Lancet Psychiatry*, *4*(9), 725–732. https://doi.org/10.1016/S2215-0366(17)30153-0

Gomes, K., & Pascual-Leone, A. (2009). Primed for Change: Facilitating Factors in Problem Gambling Treatment. *Journal of Gambling Studies*, *25*(1), 1–17. https://doi.org/10.1007/s10899-008-9111-y

Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). CONSORT-SPI 2018 Explanation and Elaboration: Guidance for reporting social and psychological intervention trials. *Trials*, *19*(1), 406. https://doi.org/10.1186/s13063-018-2735-z

Greene, C. J., Morland, L. A., Durkalski, V. L., & Frueh, B. C. (2008). Noninferiority and equivalence designs: Issues and implications for mental health research. *Journal of Traumatic Stress*, *21*(5), 433–439. https://doi.org/10.1002/jts.20367

Guidi, J., Brakemeier, E.-L., Bockting, C. L. H., Cosci, F., Cuijpers, P., Jarrett, R. B., … Fava, G. A. (2018). Methodological Recommendations for Trials of Psychological Interventions. *Psychotherapy and Psychosomatics*, *87*(5), 276–284. https://doi.org/10.1159/000490574

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Gunter, L., Zhu, J., & Murphy, S. (2007). Variable Selection for Optimal Decision Making. In R. Bellazzi, A. Abu-Hanna, & J. Hunter (Eds.), *Artificial Intelligence in Medicine* (pp. 149–154). Springer Berlin Heidelberg.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer.

Hayes, S. C., Hofmann, S. G., Stanton, C. E., Carpenter, J. K., Sanford, B. T., Curtiss, J. E., & Ciarrochi, J. (2018). The role of the individual in the coming era of process-based therapy. *Behaviour Research and Therapy*. https://doi.org/10.1016/j.brat.2018.10.005

Håkansson, A., Karlsson, A., & Widinghoff, C. (2018). Primary and Secondary Diagnoses of Gambling Disorder and Psychiatric Comorbidity in the Swedish Health Care System—A Nationwide Register Study. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/fpsyt.2018.00426

Håkansson, A., Mårdhed, E., & Zaar, M. (2017). Who Seeks Treatment When Medicine Opens the Door to Pathological Gambling Patients—Psychiatric Comorbidity and Heavy Predominance of Online Gambling. *Frontiers in Psychiatry*, *8*. https://doi.org/10.3389/fpsyt.2017.00255

He, H., Tang, W., Wang, W., & Crits-Christoph, P. (2014). Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry*, *26*(4), 236–242. https://doi.org/10.3969/j.issn.1002-0829.2014.04.008

Hedeker, D., & Gibbons, R. D. (1997). Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies. *Psychological Methods*, *2*(1), 64–78. https://doi.org/10.1037/1082-989X.2.1.64

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis* (Vol. 451). John Wiley & Sons.

Hengartner, M. P. (2018). Raising Awareness for the Replication Crisis in Clinical Psychology by Focusing on Inconsistencies in Psychotherapy Research: How Much Can We Rely on Published Findings from Efficacy Trials? *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00256

Hernán, M. A., & Hernández-Díaz, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials: Journal of the Society for Clinical Trials*, *9*(1), 48–55. https://doi.org/10.1177/1740774511420743

Hill, C. E. (2006). Introduction to special section on therapist effects. *Psychotherapy Research*, *16*(2), 143–143. https://doi.org/10.1080/10503300500470583

Hing, N., Tiyce, M., Holdsworth, L., & Nuske, E. (2013). All in the Family: Help-Seeking by Significant Others of Problem Gamblers. *International Journal of Mental Health and Addiction*, *11*(3), 396–408. https://doi.org/10.1007/s11469-012-9423-0

Hodgins, D. C., Currie, S. R., & el-Guebaly, N. (2001). Motivational enhancement and self-help treatments for problem gambling. *Journal of Consulting and Clinical Psychology*, *69*(1), 50–57. https://doi.org/10.1037//0022-006X.69.1.50

Hodgins, D. C., & el-Guebaly, N. (2000). Natural and treatment-assisted recovery from gambling problems: A comparison of resolved and active gamblers. *Addiction*, *95*(5), 777–789. https://doi.org/10.1046/j.1360-0443.2000.95577713.x

Hodgins, D. C., Stea, J. N., & Grant, J. E. (2011). Gambling disorders. *The Lancet*, *378*(9806), 1874–1884. https://doi.org/10.1016/S0140-6736(10)62185-X

Hodgins, D. C., Toneatto, T., Makarchuk, K., Skinner, W., & Vincent, S. (2007). Minimal Treatment Approaches for Concerned Significant Others of Problem Gamblers: A Randomized Controlled Trial. *Journal of Gambling Studies*, *23*(2), 215–230. https://doi.org/10.1007/s10899-006-9052-2

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1

Hofmann, S. G., Curtiss, J., & McNally, R. J. (2016). A Complex Network Perspective on Clinical Science. *Perspectives on Psychological Science*, *11*(5), 597–605. https://doi.org/10.1177/1745691616639283

Hofmann, S. G., & Hayes, S. C. (2018). The Future of Intervention Science: Process-Based Therapy. *Clinical Psychological Science*, 2167702618772296. https://doi.org/10.1177/2167702618772296

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064

Hollon, S. D. (1999). Allegiance Effects in Treatment Research: A Commentary. *Clinical Psychology: Science and Practice*, *6*(1), 107–112. https://doi.org/10.1093/clipsy.6.1.107

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., … Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, *5*(3), 237–286. https://doi.org/10.1016/S2215-0366(17)30513-8

Huhn, M., Tardy, M., Spineli, L. M., Kissling, W., Förstl, H., Pitschel-Walz, G., … Leucht, S. (2014). Efficacy of Pharmacotherapy and Psychotherapy for Adult Psychiatric Disorders: A Systematic Overview of Meta-analyses. *JAMA Psychiatry*, *71*(6), 706–715. https://doi.org/10.1001/jamapsychiatry.2014.112

Humphreys, B. R., Lee, Y. S., & Soebbing, B. P. (2010). Consumer behaviour in lottery: The double hurdle approach and zeros in gambling survey data. *International Gambling Studies*, *10*(2), 165–176. https://doi.org/10.1080/14459795.2010.502180

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 5–51. https://doi.org/10.1111/j.1467-985X.2012.01032.x

Ingle, P. J., Marotta, J., McMillan, G., & Wisdom, J. P. (2008). Significant Others and Gambling Treatment Outcomes. *Journal of Gambling Studies*, *24*(3), 381–392. https://doi.org/10.1007/s10899-008-9092-x

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision

medicine for psychiatry. *The American Journal of Psychiatry*, *171*(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8). https://doi.org/10.1371/journal.pmed.0020124

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300–307. https://doi.org/10.1037/0022-006X.67.3.300

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014). An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*, *10*(1), 99–121. https://doi.org/10.1515/ijb-2012-0052

Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., & Heagerty, P. J. (2015). The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment. *JNCI: Journal of the National Cancer Institute*, *107*(8). https://doi.org/10.1093/jnci/djv157

Jaunky, V. C., & Ramchurn, B. (2014). Consumer behaviour in the scratch card market: A double-hurdle approach. *International Gambling Studies*, *14*(1), 96–114. https://doi.org/10.1080/14459795.2013.855251

Judd, C. M., & Kenny, D. A. (1981). Process Analysis: Estimating Mediation in Treatment Evaluations. *Evaluation Review*, *5*(5), 602–619. https://doi.org/10.1177/0193841X8100500502

Julious, S. A. (2004). Sample sizes for clinical trials with Normal data. *Statistics in Medicine*, *23*(12), 1921–1986. https://doi.org/10.1002/sim.1783

Kahan, B. C., & Morris, T. P. (2013). Analysis of multicentre trials with continuous outcomes: When and how should we account for centre effects? *Statistics in Medicine*, *32*(7), 1136–1149.

Kalischuk, R. G., Nowatzki, N., Cardwell, K., Klein, K., & Solowoniuk, J. (2006). Problem gambling and its impact on families: A literature review. *International Gambling Studies*, *6*(1), 31–60. https://doi.org/10.1080/14459790600644176

Karlsson, A., & Håkansson, A. (2018). Gambling disorder, increased mortality, suicidality, and associated comorbidity: A longitudinal nationwide register study. *Journal of Behavioral Addictions*, 1–9. https://doi.org/10.1556/2006.7.2018.112

Kazdin, A. E. (2007). Mediators and Mechanisms of Change in Psychotherapy Research. *Annual Review of Clinical Psychology*, *3*(1), 1–27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

Kazdin, A. E. (2011). Evidence-based treatment research: Advances, limitations, and next steps. *American Psychologist*, *66*(8), 685. https://doi.org/10.1037/a0024975

Kazdin, A. E., & Blase, S. L. (2011). Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(1), 21–37. https://doi.org/10.1177/1745691610393527

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 285–299. https://doi.org/10.1037/0022-006X.67.3.285

Kenny, D. A. (2008). Reflections on Mediation. *Organizational Research Methods*, *11*(2), 353–358. https://doi.org/10.1177/1094428107308978

Kent, D. M., Steyerberg, E., & Klaveren, D. van. (2018). Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ*, *363*, k4245. https://doi.org/10.1136/bmj.k4245

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Khan, A., Faucett, J., Lichtenberg, P., Kirsch, I., & Brown, W. A. (2012). A Systematic Review of Comparative Efficacy of Treatments and Controls for Depression. *PLOS ONE*, *7*(7), e41778. https://doi.org/10.1371/journal.pone.0041778

Kiesler, D. J. (1966). Some myths of psychotherapy research and the search for a paradigm. *Psychological Bulletin*, *65*(2), 110.

Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the national institute of mental health treatment of depression collaborative research program data. *Psychotherapy Research*, *16*(02), 161–172.

King, K. M., Pullmann, M. D., Lyon, A. R., Dorsey, S., & Lewis, C. C. (2019). Using implementation science to close the gap between the optimal and typical practice of quantitative methods in clinical science. *Journal of Abnormal Psychology*, *128*(6), 547–562. https://doi.org/10.1037/abn0000417

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA, US: Sage Publications, Inc.

Kirsch, I. (2005). Placebo psychotherapy: Synonym or oxymoron? *Journal of Clinical Psychology*, *61*(7), 791–803. https://doi.org/10.1002/jclp.20126

Kirsch, I., Wampold, B. E., & Kelley, J. M. (2016). Controlling for the placebo effect in psychotherapy: Noble quest or tilting at windmills? *Psychology of Consciousness: Theory, Research, and Practice*, *3*(2), 121–131. https://doi.org/10.1037/cns0000065

Krishnan, M., & Orford, J. (2002). Gambling and the family: From the stress-coping-support Perspective. *International Gambling Studies*, *2*(1), 61–83. https://doi.org/10.1080/14459790208732300

Lambert, M. J. (1992). Psychotherapy outcome research: Implications for integrative and eclectical therapists. In *Handbook of psychotherapy integration* (pp. 94–129). New York, NY, US: Basic Books.

Lambert, M. J. (Ed.). (2013). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.). John Wiley & Sons.

Langham, E., Thorne, H., Browne, M., Donaldson, P., Rose, J., & Rockloff, M. (2016). Understanding gambling related harm: A proposed definition, conceptual framework, and taxonomy of harms. *BMC Public Health*, *16*(1), 80. https://doi.org/10.1186/s12889-016-2747-0

Lappalainen, P., Granlund, A., Siltanen, S., Ahonen, S., Vitikainen, M., Tolvanen, A., & Lappalainen, R. (2014). ACT Internet-based vs face-to-face? A randomized controlled trial of two ways to deliver Acceptance and Commitment Therapy for depressive symptoms: An 18-month follow-up. *Behaviour Research and Therapy*, *61*, 43–54. https://doi.org/10.1016/j.brat.2014.07.006

Leichsenring, F., Abbass, A., Driessen, E., Hilsenroth, M., Luyten, P., Rabung, S., & Steinert, C. (2018). Equivalence and non-inferiority testing in psychotherapy research. *Psychological Medicine*, *48*(11), 1917–1919. https://doi.org/10.1017/S0033291718001289

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., … Steinert, C. (2017). Biases in research: Risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Leykin, Y., & DeRubeis, R. J. (2009). Allegiance in Psychotherapy Outcome Research: Separating Association From Bias. *Clinical Psychology: Science and Practice*, *16*(1), 54–65. https://doi.org/10.1111/j.1468-2850.2009.01143.x

Lieb, K., Osten-Sacken, J. von der, Stoffers-Winterling, J., Reiss, N., & Barth, J. (2016). Conflicts of interest and spin in reviews of psychological therapies: A systematic review. *BMJ Open*, *6*(4), e010606. https://doi.org/10.1136/bmjopen-2015-010606

Lipkovich, I., Dmitrienko, A., & D'Agostino, R. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136–196. https://doi.org/10.1002/sim.7064

Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121. https://doi.org/10.1080/01621459.1995.10476615

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.

Lu, W., Zhang, H. H., & Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, *22*(5), 493–504. https://doi.org/10.1177/0962280211428383

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., … Schweizer, E. (1999). The Researcher's Own Therapy Allegiances: A "Wild Card" in Comparisons of Treatment Efficacy. *Clinical Psychology: Science and Practice*, *6*(1), 95–106. https://doi.org/10.1093/clipsy.6.1.95

Luedtke, A. R., & van der Laan, M. J. (2017). Evaluating the impact of treating the optimal subgroup. *Statistical Methods in Medical Research*, *26*(4), 1630–1640. https://doi.org/10.1177/0962280217708664

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., … Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, *383*(9912), 101–104. https://doi.org/10.1016/S0140-6736(13)62329-6

Madden, D. (2008). Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, *27*(2), 300–307. https://doi.org/10.1016/j.jhealeco.2007.07.001

Magnusson, K. (2018). Technical Appendix: Details on the Power Calculations for Two-and Three-level Models with Missing Data. Retrieved from https://cran.r-project.org/web/packages/powerlmm/vignettes/technical.pdf

Magnusson, K., Nilsson, A., Hellner Gumpert, C., Andersson, G., & Carlbring, P. (2015). Internet-delivered cognitive-behavioural therapy for concerned significant others of people with problem gambling: Study protocol for a randomised wait-list controlled trial. *BMJ Open*, *5*(12), e008724. https://doi.org/10.1136/bmjopen-2015-008724

Makarchuk, K., Hodgins, D. C., & Peden, N. (2002). Development of a Brief Intervention for Concerned Significant Others of Problem Gamblers. *Addictive Disorders & Their Treatment*, *1*(4), 126–134. https://doi.org/10.1097/00132576-200211000-00003

Martindale, C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *46*(6), 1526–1530. https://doi.org/10.1037/0022-006X.46.6.1526

Meehl, P. E. (1955). Psychotherapy. *Annual Review of Psychology*, *6*(1), 357–378. https://doi.org/10.1146/annurev.ps.06.020155.002041

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Meis, L. A., Griffin, J. M., Greer, N., Jensen, A. C., MacDonald, R., Carlyle, M., … Wilt, T. J. (2013). Couple and family involvement in adult mental health treatment: A systematic review. *Clinical Psychology Review*, *33*(2), 275–286. https://doi.org/10.1016/j.cpr.2012.12.003

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, *20*(8), 897–916. https://doi.org/10.1002/hec.1653

Min, Y., & Agresti, A. (2002). Modeling Nonnegative Data with Clumping at Zero: A Survey. *Journal of the Iranian Statistical Society*, *1*(1), 7–33. Retrieved from http://jirss.irstat.ir/article-1-84-en.html

Mohammad Maracy, & Graham Dunn. (2011). Estimating dose-response effects in psychological treatment trials: The role of instrumental variables. *Statistical Methods in Medical Research*, *20*(3), 191–215. https://doi.org/10.1177/0962280208097243

Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Arean, P., Hollon, S. D., … Kaplan, R. (2009). The Selection and Design of Control Conditions for Randomized Controlled Trials of Psychological Interventions. *Psychotherapy and Psychosomatics*, *78*(5), 275–284. https://doi.org/10.1159/000228248

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 371–388. https://doi.org/10.1111/j.1467-9868.2007.00640.x

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107587991

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *0*(0). https://doi.org/10.1002/sim.8086

Mulder, R., Murray, G., & Rucklidge, J. (2017). Common versus specific factors in

psychotherapy: Opening the black box. *The Lancet Psychiatry*, *4*(12), 953–962. https://doi.org/10.1016/S2215-0366(17)30100-1

Munder, T., Brütsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review*, *33*(4), 501–511. https://doi.org/10.1016/j.cpr.2013.02.002

Nathan, P. E., & Gorman, J. M. (2015). *A guide to treatments that work*. Oxford University Press.

Nayoski, N., & Hodgins, D. C. (2016). The Efficacy of Individual Community Reinforcement and Family Training (CRAFT) for Concerned Significant Others of Problem Gamblers. *Journal of Gambling Issues*, (33), 189. https://doi.org/10.4309/jgi.2016.33.11

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research Part 1: Background and overview. *Statistics in Medicine*, *35*(27), 5070–5093. https://doi.org/10.1002/sim.7050

Newman, S. C., & Thompson, A. H. (2007). The Association between Pathological Gambling and Attempted Suicide: Findings from a National Survey in Canada. *The Canadian Journal of Psychiatry*, *52*(9), 605–612. https://doi.org/10.1177/070674370705200909

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241.

Nilsson, A., Magnusson, K., Carlbring, P., Andersson, G., & Hellner Gumpert, C. (2016). Effects of added involvement from concerned significant others in internet-delivered CBT treatments for problem gambling: Study protocol for a randomised controlled trial. *BMJ Open*, *6*(9), e011974. https://doi.org/10.1136/bmjopen-2016-011974

Nock, M. K. (2007). Conceptual and Design Essentials for Evaluating Mechanisms of Change. *Alcoholism: Clinical and Experimental Research*, *31*(s3), 4s–12s. https://doi.org/10.1111/j.1530-0277.2007.00488.x

Norcross, J. C., Beutler, L. E., & Levant, R. F. (2006). *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions.* American Psychological Association.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. Retrieved from http://arxiv.org/abs/1205.4251

Ogles, B. M. (2013). Measuring change in psychotherapy research. In *Bergin and*

Garfield's handbook of psychotherapy and behavior change (Vol. 6). Hoboken, New Jersey: John Wiley & Sons.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*(454), 730–745. https://doi.org/10.1198/016214501753168389

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Os, J. van, Guloksuz, S., Vijn, T. W., Hafkenscheid, A., & Delespaul, P. (2019). The evidence-based group-level symptom-reduction model as the organizing principle for mental health care: Time for change? *World Psychiatry*, *18*(1), 88–96. https://doi.org/10.1002/wps.20609

Owen, J., Drinane, J. M., Idigo, K. C., & Valentine, J. C. (2015). Psychotherapist effects in meta-analyses: How accurate are treatment effects? *Psychotherapy (Chicago, Ill.)*, *52*(3), 321–328. https://doi.org/10.1037/pst0000014

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118.

Petry, N. M., Weinstock, J., Ledgerwood, D. M., & Morasco, B. (2008). A randomized trial of brief interventions for problem and pathological gamblers. *Journal of Consulting and Clinical Psychology*, *76*(2), 318–328. https://doi.org/10.1037/0022-006X.76.2.318

Petry, N. M., & Weiss, L. M. (2009). Social support is associated with gambling treatment outcomes in pathological gamblers. *The American Journal on Addictions / American Academy of Psychiatrists in Alcoholism and Addictions*, *18*(5), 402–408. https://doi.org/10.3109/10550490903077861

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J. W., & Group, for the C. (2006). Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement. *JAMA*, *295*(10), 1152–1160. https://doi.org/10.1001/jama.295.10.1152

Preacher, K. J. (2015). Advances in Mediation Analysis: A Survey and Synthesis of New Developments. *Annual Review of Psychology*, *66*(1), 825–852. https://doi.org/10.1146/annurev-psych-010814-015258

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Rhoads, C. H. (2012). Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies. *Statistics, Politics, and Policy*, *3*(1). https://doi.org/10.1515/2151-7509.1012

Rief, W., & Hofmann, S. G. (2018). Some problems with non-inferiority tests in psychotherapy research: Psychodynamic therapies as an example. *Psychological Medicine*, *48*(8), 1392–1394. https://doi.org/10.1017/S0033291718000247

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*(2), 152–162.

Roozen, H. G., Waart, R. D., & Kroft, P. V. D. (2010). Community reinforcement and family training: An effective option to engage treatment-resistant substance-abusing individuals in treatment. *Addiction*, *105*(10), 1729–1738. https://doi.org/10.1111/j.1360-0443.2010.03016.x

Rosenzweig, S. (1936). Some Implicit Common Factors in Diverse Methods of Psychotherapy. *American Journal of Orthopsychiatry*, *6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rude, J., Surry, Y., & Kron, R. (2014). A generalized double-hurdle model of Swedish gambling expenditures. *Applied Economics*, *46*(34), 4151–4163. https://doi.org/10.1080/00036846.2014.939376

Rychtarik, R. G., & McGillicuddy, N. B. (2006). Preliminary Evaluation of a Coping Skills Training Program for Those with a Pathological-Gambling Partner. *Journal of Gambling Studies*, *22*(2), 165–178. https://doi.org/10.1007/s10899-006-9008-6

Sawkins, J. W., & Dickie, V. A. (2002). National Lottery participation and expenditure: Preliminary results using a two stage modelling approach. *Applied Economics Letters*, *9*(12), 769–773. https://doi.org/10.1080/13504850210129441

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org//1082-989X.7.2.147

Schielzeth, H., & Nakagawa, S. (2013). Nested by design: Model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, *4*(1), 14–24. https://doi.org/10.1111/j.2041-210x.2012.00251.x

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood

estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.

Senn, S. (2004). Individual response to treatment: Is it a valid assumption? *BMJ*, *329*(7472), 966–968. https://doi.org/10.1136/bmj.329.7472.966

Senn, S. (2016). Mastering variation: Variance components and personalised medicine. *Statistics in Medicine*, *35*(7), 966–977. https://doi.org/10.1002/sim.6739

Siemer, M., & Joormann, J. (2003). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods*, *8*(4), 497–517. https://doi.org/10.1037/1082-989X.8.4.497

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, V. A., Neelon, B., Preisser, J. S., & Maciejewski, M. L. (2015). A marginalized two-part model for longitudinal semicontinuous data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280215592908

Snijders, T. (2005). Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science*.

Statens folkhälsoinstitut. (2010). Spel om pengar och spelproblem i Sverige 2008/2009: Huvudresultat från SWELOGS befolkningsstudie.

Steinert, C., Munder, T., Rabung, S., Hoyer, J., & Leichsenring, F. (2017). Psychodynamic Therapy: As Efficacious as Other Empirically Supported Treatments? A Meta-Analysis Testing Equivalence of Outcomes. *American Journal of Psychiatry*, *174*(10), 943–953. https://doi.org/10.1176/appi.ajp.2017.17010057

Stern, S. E., & Welsh, A. H. (2000). Likelihood inference for small variance components. *Canadian Journal of Statistics*, *28*(3), 517–532.

Steyerberg, E. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer-Verlag.

Steyerberg, E. W., Uno, H., Ioannidis, J. P. A., Calster, B. van, Ukaegbu, C., Dhingra, T., … Kastrinos, F. (2018). Poor performance of clinical prediction models: The harm of commonly applied methods. *Journal of Clinical Epidemiology*, *98*, 133–143. https://doi.org/10.1016/j.jclinepi.2017.11.013

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge,*

Mass.), *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 1171–1177.

Stranahan, H. A., & Borg, M. O. (1998). Separating the Decisions of Lottery Expenditures and Participation: A Truncated Tobit Approach. *Public Finance Review*, *26*(2), 99–117. https://doi.org/10.1177/109114219802600201

Su, L., Tom, B. D. M., & Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, *10*(2), 374–389. https://doi.org/10.1093/biostatistics/kxn044

Svensson, J., Romild, U., & Shepherdson, E. (2013). The concerned significant others of people with gambling problems in a national representative sample in Sweden – a 1 year follow-up study. *BMC Public Health*, *13*(1), 1087. https://doi.org/10.1186/1471-2458-13-1087

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., … Shrout, P. E. (2017). It's Time to Broaden the Replicability Conversation: Thoughts for and From Clinical Psychological Science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Toh, S., & Hernán, M. A. (2008). Causal Inference from Longitudinal Studies with Baseline Randomization. *The International Journal of Biostatistics*, *4*(1). https://doi.org/10.2202/1557-4679.1117

Toneatto, T., & Ladoceur, R. (2003). Treatment of pathological gambling: A critical review of the literature. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *17*(4), 284–292. https://doi.org/10.1037/0893-164X.17.4.284

Tooze, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, *11*(4), 341–355. Retrieved from http://smm.sagepub.com/content/11/4/341.short

Valente, M. J., Pelham, W. E., Smyth, H., & MacKinnon, D. P. (2017). Confounding in statistical mediation analysis: What it is and how to address it. *Journal of Counseling Psychology*, *64*(6), 659–671. https://doi.org/10.1037/cou0000242

Valentine, G., & Hughes, K. (2010). Ripples in a pond: The disclosure to, and management of, problem Internet gambling with/in the family. *Community, Work & Family*, *13*(3), 273–290. https://doi.org/10.1080/13668803.2010.488107

van Klaveren, D., Steyerberg, E. W., Serruys, P. W., & Kent, D. M. (2018). The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *Journal of Clinical Epidemiology*, *94*, 59–68.

https://doi.org/10.1016/j.jclinepi.2017.10.021

Volberg, R. A., Abbott, M. W., Rönnberg, S., & Munck, I. M. (2001). Prevalence and risks of pathological gambling in Sweden. *Acta Psychiatrica Scandinavica*, *104*(4), 250–256. https://doi.org/10.1034/j.1600-0447.2001.00336.x

Walker, M., Toneatto, T., Potenza, M. N., Petry, N., Ladouceur, R., Hodgins, D. C., … Blaszczynski, A. (2006). A framework for reporting outcomes in problem gambling treatment research: The Banff, Alberta Consensus. *Addiction*, *101*(4), 504–511. https://doi.org/10.1111/j.1360-0443.2005.01341.x

Wallach, J. D., Boyack, K. W., & Ioannidis, J. P. A. (2018). Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biology*, *16*(11), e2006930. https://doi.org/10.1371/journal.pbio.2006930

Walwyn, R., & Roberts, C. (2010). Therapist variation within randomised trials of psychotherapy: Implications for precision, internal and external validity. *Statistical Methods in Medical Research*, *19*(3), 291–315. https://doi.org/10.1177/0962280209105017

Walwyn, R., & Roberts, C. (2015). Meta-analysis of absolute mean differences from randomised trials with treatment-related clustering associated with care providers. *Statistics in Medicine*, *34*(6), 966–983.

Wampold, B. E. (2005). Establishing Specificity in Psychotherapy Scientifically: Design and Evidence Issues. *Clinical Psychology: Science and Practice*, *12*(2), 194–197. https://doi.org/10.1093/clipsy.bpi025

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wampold, B. E., Frost, N. D., & Yulish, N. E. (2016). Placebo effects in psychotherapy: A flawed concept and a contorted history. *Psychology of Consciousness: Theory, Research, and Practice*, *3*(2), 108–120. https://doi.org/10.1037/cns0000045

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, *5*(4), 425.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC. https://doi.org/10.1201/EBK1439808184

Wenzel, H. G., Øren, A., & Bakken, I. J. (2008). Gambling problems in the family – A

stratified probability sample study of prevalence and reported consequences. *BMC Public Health*, *8*(1), 412. https://doi.org/10.1186/1471-2458-8-412

Westphal, J. R. (2007). Are the Effects of Gambling Treatment Overestimated? *International Journal of Mental Health and Addiction*, *5*(1), 65–79. https://doi.org/10.1007/s11469-006-9038-4

Westphal, J. R. (2008). How Well are We Helping Problem Gamblers? An Update to the Evidence Base Supporting Problem Gambling Treatment. *International Journal of Mental Health and Addiction*, *6*(2), 249–264. https://doi.org/10.1007/s11469-007-9072-x

Williams, R. J., Volberg, R. A., & Stevens, R. M. (2012). *The population prevalence of problem gambling: Methodological influences, standardized rates, jurisdictional differences, and worldwide trends*. Ontario Problem Gambling Research Centre.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Xing, D., Huang, Y., Chen, H., Zhu, Y., Dagne, G. A., & Baldwin, J. (2015). Bayesian inference for two-part mixed-effects model using skew distributions, with application to longitudinal semicontinuous alcohol data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280215590284

Zhang, B., Liu, W., & Hu, Y. (2018). Estimating marginal and incremental effects in the analysis of medical expenditure panel data using marginalized two-part random-effects generalized Gamma models: Evidence from China healthcare cost data. *Statistical Methods in Medical Research*, *27*(10), 3039–3061. https://doi.org/10.1177/0962280217690770