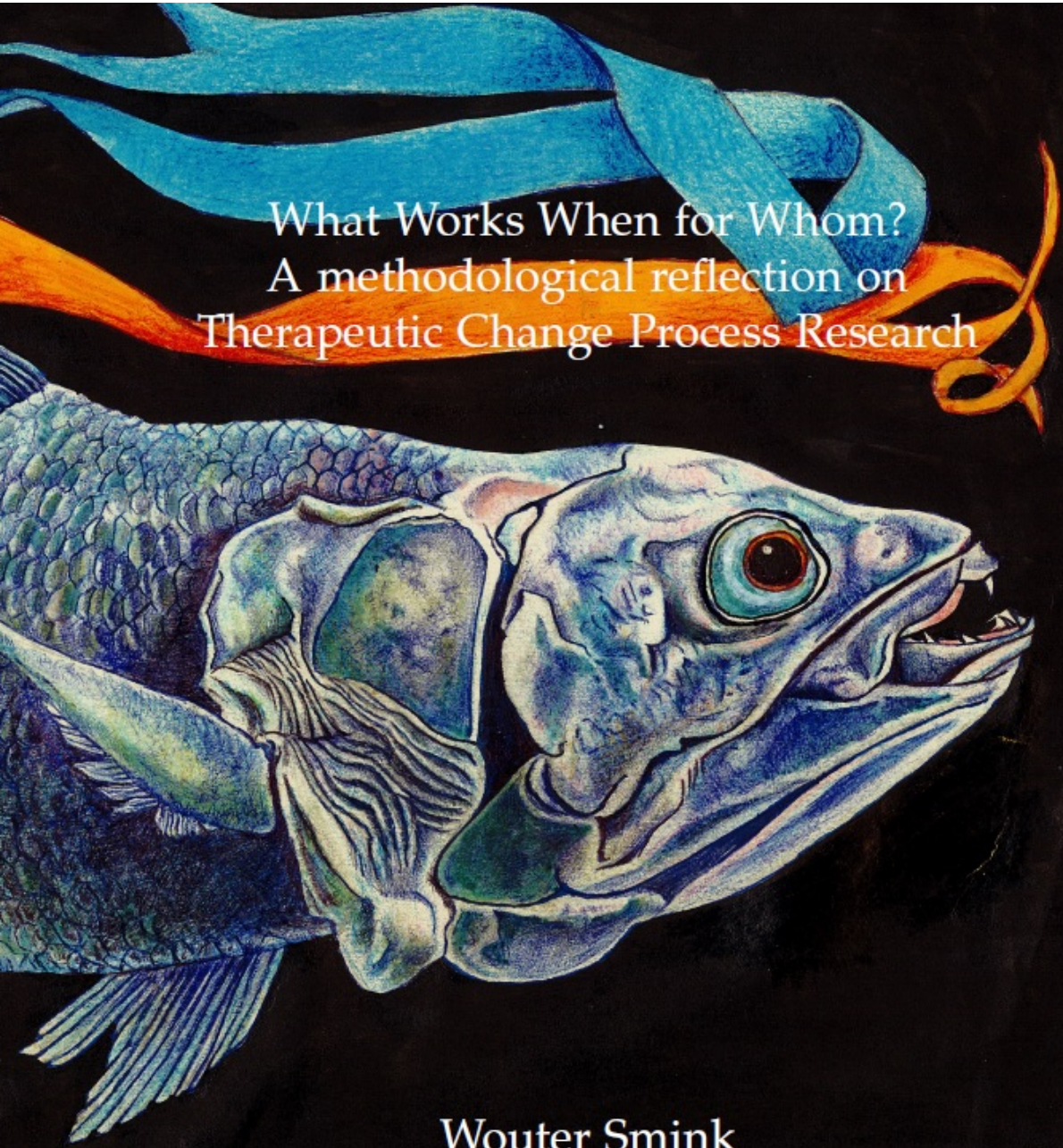
What Works When for Whom? A methodological reflection on Therapeutic Change Process Research



Wouter Smink

What Works When for Whom?
A methodological reflection on
Therapeutic Change Process Research

Wouter Smink

This thesis is part of the *What Works When for Whom* project, supported by the Life Science & eHealth domain of the Accelerating Scientific Discovery (ASDI) call from the Netherlands eScience Center (NLeSC; Amsterdam, the Netherlands): grant number 027.015.G04 awarded to dr. A. M. Sools. The NLeSC is the national knowledge center for the development and application of research software to advance scientific research, and is funded by the Netherlands Organization for Scientific Research (in *Dutch*: *Nederlandse organisatie voor Wetenschappelijk Onderzoek*) and SURF (*Samenwerkende Universitaire Rekenfaciliteiten*). The project received no additional funding, which allowed us to conduct all of the thesis' research projects in the absence of any conflicting interests. Printing was financially supported by the University of Twente.

Design	Rob Smink ("*na een volledige lezing zullen alle tekeningen duidelijk zijn*")
Printed by	Gildeprint Drukkerijen in Enschede
Typesetting	LaTeX (edited in TeXstudio)
References	APA 7 through BibTeX and Mendeley
ISBN	978-90-365-5033-8
DOI	10.3990/1.9789036550338

What Works When for Whom?

A methodological reflection on *Therapeutic Change Process Research*

PROEFSCHRIFT

(met een samenvatting in het Nederlands)

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de Rector Magnificus,
Prof. dr. ir. A. Veldkamp,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 12 februari 2021 om 14.45 uur
in de Prof. dr. G. Berkhoff-Zaal in *De Waaier*
(gebouw nr. 12 op de Universiteit Twente)

door

Wouter Arend Christiaan Smink

geboren op 1 juli 1992
in Hengelo (Overijssel)

Dit proefschrift is goedgekeurd door:

de promotoren

Prof.	dr.	G. J. Westerhof	Universiteit Twente
Prof.	dr. ir.	B. P. Veldkamp	Universiteit Twente

de co-promotor

	dr.	A. M. Sools	Universiteit Twente

De openbare verdediging van dit proefschrift vindt plaats ten overstaan van de **promotiecommissie** die bestaat uit:

de voorzitter / secretaris

Prof.	dr.	T. A. J. Toonen	Universiteit Twente

de promotoren

Prof.	dr.	G. J. Westerhof	Universiteit Twente
Prof.	dr. ir.	B. P. Veldkamp	Universiteit Twente

de co-promoter

	dr.	A. M. Sools	Universiteit Twente

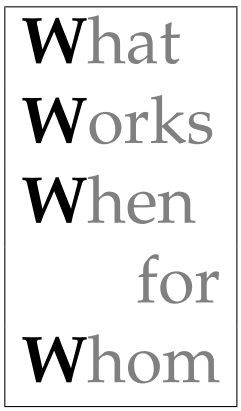
en **de leden**

Prof.	dr.	E. T. Bohlmeijer	Universiteit Twente
Prof.	dr.	D. K. J. Heylen	Universiteit Twente
Prof.	dr.	I. G. Klugkist	Universiteit Utrecht
	dr.	J. E. L. van der Nagel	Tactus Verslavingzorg, Universiteit Twente
Prof.	dr.	H. Riper	Vrije Universiteit Amsterdam

Begeleiding tijdens de verdediging wordt gegeven door:

de paranimfen

	drs.	A. D. Berkien	oud-docent Twickel College
	drs.	P. D. Noort	Universiteit Twente

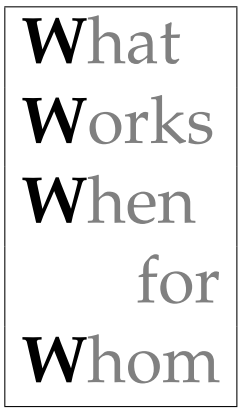

What
Works
When
for
Whom

Abstract

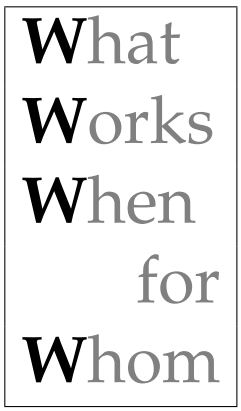
We aim to advance *Therapeutic Change Process Research* (TCPR), a field dedicated to find out *what* treatment –by *whom* and under *which* set of circumstances– is most effective for *this* individual with *that* specific problem. Our approach advocates that assessing the therapeutic exchange between client and counsellor provides a possibility to open the 'black box' of therapy to learn more about *What Works When for Whom* (WWWW). Web-based interventions provide an unique opportunity for TCPR: as online counselling is effective, all active ingredients of therapy should be included in the exchanged e-mails. Through seven propositions, we argue why the e-mail based 'talking cure' contains a wealth of information about the WWWW question, and present an approach that consists out of three parts. In the first part of the thesis, we discuss the automated and qualitative TCPR *methods* that are used to study language. In the second, we discuss the TCPR *models* that are (and should be) used to model the results of these methods. We reflect on the differences between the models and methods through the *automation-explication* framework. We favour multilevel modelling methods for TCPR, but these models have a shortcoming: they cannot assess negative clustering effects. In the last part, we present a gentle introduction to *Bayesian Covariance Structure Modelling*: an *alternative* TCPR *model* that is capable of addressing the WWWW question by modelling negative clustering effects.

Nederlandse vertaling van het abstract

We beogen om het veld van *Therapeutic Change Process Research* (TCPR; vrij vertaald: 'onderzoek naar therapeutische veranderprocessen') voortuit te helpen. Dit veld is gedreven door de vraag *Wat* [voor psychotherapie] *Werkt er Wanneer en voor Wie* (WWWW; *What Works When for Whom*)? Onze aanpak bestaat uit het bestuderen van de therapeutische interactie tussen client en counsellor om zo de therapeutisch 'black box' te openen in antwoord op de WWWWvraag. Aan de hand van zeven stellingen beargumenteren we dat veel van de informatie die relevant is voor TCPR moet zitten in de e-mails die worden uitgewisseld in online interventies. Onze aanpak bestaat uit drie delen die we in deze thesis uiteenzetten: in het eerste bediscussiëren we de geautomatiseerde en kwalitatieve TCPR*methoden* die gebruikt worden om taal in e-mails te bestuderen. In het tweede deel bespreken we de verschillende TCPR*modellen* die voor dit doeleinde gebruikt worden. We reflecteren op het verschil in de methoden en modellen aan de hand van het *automation-explication* framework. We hebben een voorkeur voor de (uitlegbare) multilevel modellen, maar deze modellen hebben ook een tekortkoming: ze kunnen niet het negatieve (vrij vertaald: 'divergente') effect van clusters modelleren. In het laatste deel van de thesis besteden we hier aandacht aan en introduceren we *Bayesian Covariance Structure Modelling*. Dit *alternatieve* TCPR*model* beantwoordt de WWWWvraag door juist deze divergente effecten te modelleren.

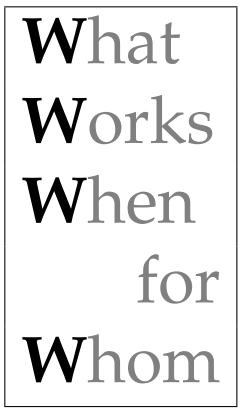

What
Works
When
for
Whom

Chapters


What
Works
When
for
Whom

Often-used abbreviations.

TCPR	Therapeutic Change Process Research
WWWW	What Works When for Whom
LIWC	Linguistic Inquiry and Word Count (pronounced as the name '*Luke*') a program by Pennebaker, Boyd, et al. (2015)
AdB	*Alcohol de Baas* (translated from *Dutch* as 'Look at your drinking')
AUD	Alcohol Use Disorder (in chapter 5)
MLM	Multilevel Model (in chapter 6; LME in chapter 7)
LM	Linear Model (in chapter 7)
LME	Linear Mixed Effects model (in chapter 7; MLM in chapter 6)
BCSM	Bayesian Covariance Structure Modelling (in chapter 8)


What
Works
When
for
Whom

A general introduction of the thesis

Woody Oh, what?! What?! These are plastic. He can't fly!

Buzz Lightyear They are a terillium-carbonic alloy and I *can* fly.

Woody No, you can't.

Buzz Yes, I can.

Woody You can't!

Buzz Can!

Woody Can't! Can't! Can't!

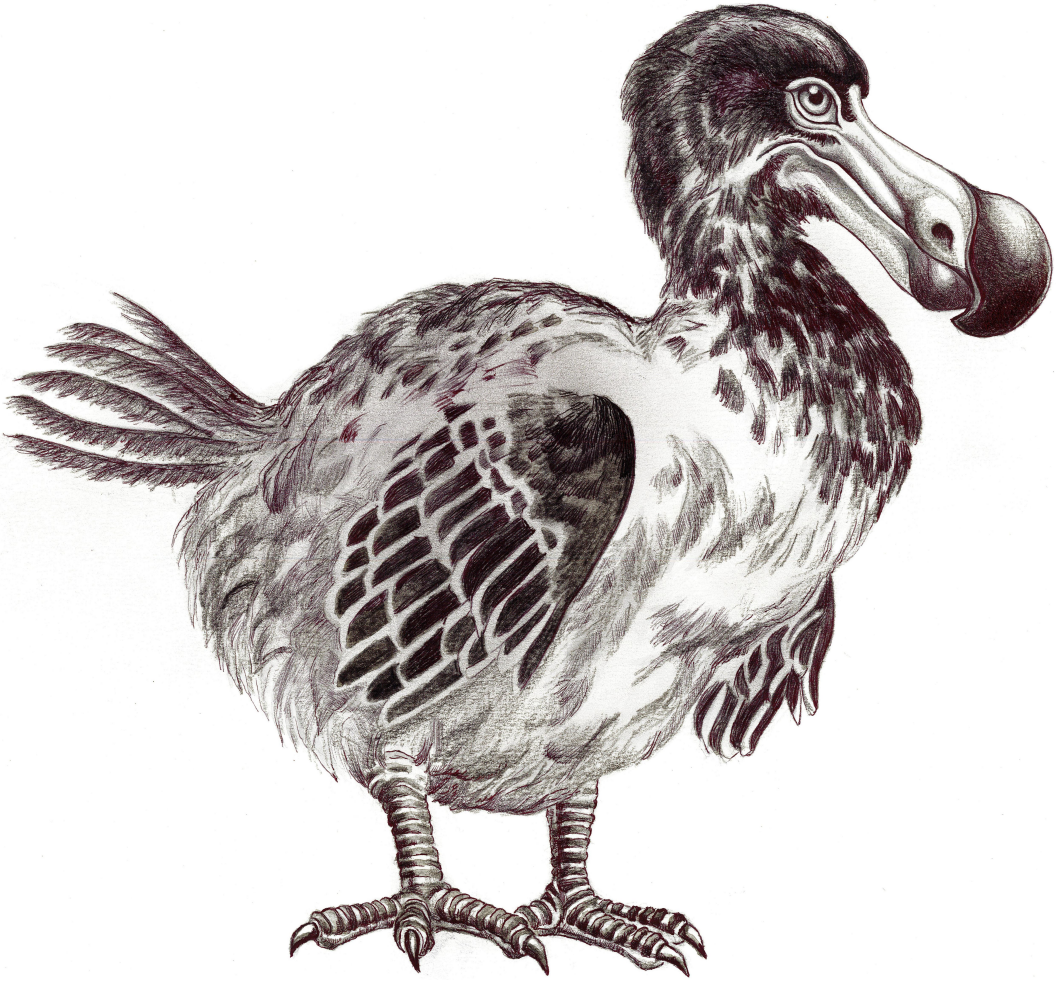
Buzz I tell you, I could fly around this room with my eyes closed!

Woody Okay then, Mr. Lightbeer! Prove it.

Buzz All right then, I will. Stand back everyone!

Woody That wasn't flying! That was falling with style!

From the movie *Toy Story* by John Lasseter (1995)
produced at Pixar Animation Studios



1 General introduction and outline

> At last the Dodo said,
> "*Everybody* has won, and all
> must have prizes."
>
> *Alice's Adventures in Wonderland*
> (1865, p. 34) by Lewis Carroll

Introduction

Mark Twain –author of *The Adventures of Huckleberry Finn* and *Tom Sawyer*– is one of the most famous American writers.[1] Twain was twenty-six years old when hostilities between the northern and southern United States broke out, and his role in the Civil War has long been the subject of dispute (Brinegar, 1963): the issue is whether he participated, and –if he did– on which side he fought (Larsen & Marx, 2012, p. 4). Twain always dodged the question, and took the answer to his grave. So ... case closed?

The short answer: no. I remember the surprise I felt in high school when I read about statistical tools for automatic text analyses, and how they could help to reassess cases like Twain's. Some historians argue that ten essays in the *New Orleans Daily Cresent* could solve the mystery (Larsen & Marx, 2012, p. 4). Written by "Quintus Curtius Snodgrass", the essays described the war through the eyes of a member of the Louisiana militia.

However, the army's archives do not contain anyone with the name 'Snodgrass'. Also, the letters display the sense of humour and irony that were typical for Twain's work (Larsen & Marx, 2012, p. 4). Did Twain –the pen name of Samual Langhorne Clemens– also use Snodgrass as a pseudonym?

A simple method to answer this question is to *count the words* of Twain and Snodgrass. The frequency by which an author chooses words form an author-specific *frequency distribution*. In other words: statistical assessment of language-

[1] The following case comes *directly* from Larsen and Marx (2012, p. 4 – 5, p. 460 – 464).

Table 1.1: The proportion of three-letter words of Mark Twain and Quintus Curtius Snodgrass.

Mark Twain	pp.	QCS	pp.
Sergeant Fathom letter	0.225	Letter I	0.209
Madame Caprell letter	0.226	Letter II	0.205
Mark Twain letters in		Letter III	0.196
Territorial Enterprise		Letter IV	0.210
First letter	0.217	Letter V	0.202
Second letter	0.240	Letter VI	0.207
Third letter	0.230	Letter VII	0.224
Fourth letter	0.229	Letter VIII	0.223
First *Innocents Abroad* letter		Letter IX	0.220
First half	0.235	Letter X	0.201
Second half	0.217		
Average pp.	0.232		0.210

Note. Reprinted from Larsen and Marx (2012, p. 5) by the permission of Pearson Education, Inc.
pp. Proportion of the number of three-letter words.
QCS Quintus Curtius Snodgrass.

use could make text characteristics –such as *word length*, *number of verbs*, and the *proportion of personal pronouns*– as unique as a fingerprint (Brinegar, 1963).

Table 1.1 displays the proportion of *three-letter words* that Twain and Snodgrass use in several of their writings. With this Table, the question of whether the authors are the same person is now a matter of choosing between one of the following hypotheses (Larsen & Marx, 2012, p. 5):

H_0: The difference between proportions is so small (i.e. close to 0) that it is reasonable to rule out possibility that Twain and Snodgrass are *different* persons (i.e. Twain = Snodgrass).

H_a: The difference between proportions is so large that the only reasonable conclusion is that Twain and Snodgrass are **not** the same person (i.e. Twain $\neq$ Snodgrass).

Based on Table 1.1 we calculated Twain's *average* proportion of three-letter words as 23.2%, and 21.0% for Snodgrass. These proportions are obviously **not** similar, but do the proportions also differ with a *degree of statistical significance*? Choosing between H_0 and H_a is –in fact– a statistical problem that can be addressed by calculating the t-statistic.

An Equation is helpful (but this paragraph can be skipped). We refer to the average proportions of Twain and Snodgrass as $\bar{x}_1$ and $\bar{x}_2$ respectively. The number of observed proportions for Twain is 8 (n_1), and 10 for Snodgrass (n_2). The pooled standard deviation s_p is 0.012.[2] The corresponding t-statistic can then be calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.231 - 0.210}{0.012\sqrt{\frac{1}{8} + \frac{1}{10}}} = 0.022. \tag{1.1}$$

A t-statistic with a value as small as the one in Equation (1.1) favours H_0.[2]

In other words, an analysis indicates that it is likely that Twain and Snodgrass are *one and the same person*! So ... case closed? The short answer is (again) *no*. Is one statistical analysis sufficient for claiming that two authors are one and the same person? Brinegar (1963) conducted a batch of analyses with other text characteristics, and almost all favoured H_a, meaning that it is –in fact– highly *unlikely* that Twain wrote the Snodgrass letters.

Profile of the thesis

Text mining offers to opportunity to systematically compare texts based on aspects that are not always straightforward to measure manually (such as the proportion of three-letter words). But –as the example shows– it leaves open what aspects of a text should be studied: it is not always straightforward how meaningful conclusions can be based on text mining research. In this thesis, we will argue that text mining questions pertaining to ***T**herapeutic **C**hange **P**rocess **R**esearch* (TCPR) require a *multidisciplinary* and *multilevelled* analysis. In the next section we address the context that makes TCPR stand out.

And as this thesis is the result of what I've learned since I first read about the Snodgrass letters, it felt natural to return to the case that introduced me to the field (and don't we all enjoy a mystery)?

Therapeutic Change Process Research

The overarching goal of the thesis is to advance TCPR. Our approach adheres to the statistical analysis of language in similar fashion to the Snodgrass-case. In the next section we discuss the research questions that we address throughout the thesis, but first we discuss the *global impact* of mental health issues.

[2]For the full calculation, see Larsen and Marx (2012, p. 460 – 464).

The global impact of mental health issues

It is well-known that mental health problems are one of the most significant causes of the global disease burden (Vos et al., 2015), with many people having some degree of experience with a mental health issue. The staggering and worldwide impact of mental illness is established through decades of research (WHO; Degenhardt et al., 2018): more than one in three people meet criteria for diagnoses of a mental disorder at some point in their life (Andrade et al., 2013; Ginn & Horder, 2012; Steel et al., 2014), so it is not surprising that one in six people experienced a mental health problem in the past week (McManus et al., 2016).

These statistics stress the worldwide impact: around 25% of the global population and about 33% of Europe have a mental health issue every year (Formánek et al., 2019), and these numbers increase annually (Andrade et al., 2013). It is likely that these numbers should be higher, as in some parts of the world it is difficult to assess mental health of people (and the available numbers differ between reports; Harpham & Molyneux, 2001; Karpenko & Kostyuk, 2018; Lin, 1983; Que et al., 2019; van Voren, 2017).

In this thesis, we will advocate a methodological approach to advancing TPCR. With this emphasis, we have to limit our scope and chose to focus on two of the most common mental health disorders.

Depression

Depression alone is among the leading causes of disability worldwide, and a major contributor to the burden of suicide and ischaemic heart disease (Whiteford et al., 2013). In the Netherlands alone, 18% of the adults between 18 and 65 years old suffered from mental illness during the last year, and around one million individuals seek psychological treatment each year (de Graaf et al., 2010). Mental health problems cause important limitations in social functioning and in quality of life, and contribute to about a quarter of the losses in Dutch health life years (de Graaf et al., 2011). There are various interventions available to treat depression, but in this thesis, we specifically focus on web-based interventions that rely on the exchange of e-mail between client and counsellor.

The 'Op Verhaal Komen'-intervention The data we use to study depression were collected as part of the *Op Verhaal Komen*-intervention ('The stories we live by' in *Dutch*; Bohlmeijer & Westerhof, 2010). Lamers et al. (2015) used a randomized controlled trial to investigate the short-term and long-term effects of

an e-mail-guided intervention. We extent their analyses by assessing the content of their e-mails through similar methods as the Snodgrass-case.

Alcohol use disorder

AUD is the most prevalent substance use disorders of all with 283 million individuals affected globally (Ball et al., 2006; Degenhardt et al., 2018; Rehm et al., 2013). Alcohol-related problems are difficult to treat: total treatment duration of AUD is –on average– up to 18 years (Bruffaerts et al., 2007; Chapman et al., 2015; Korbmacher, 2014), with only one in three problematic drinkers ever seeking treatment (Cunningham & Breslin, 2004). However, especially for AUD (Cloud & Peacock, 2001), it turns out that web-based interventions have a lower threshold than face-to-face therapy (Vernon, 2010).

The 'Alcohol de Baas'-intervention Because those who are alcohol dependent are more active online nowadays, data from web-based interventions hold potential for understanding AUD. The data we use for AUD were collected by Marloes Postel (2011), in association with Tactus *Verslavingszorg* ('addiction care' in *Dutch*). Similar to the study of Lamers et al. (2015), the content of the e-mails were not included in the original analyses, which is how we extent the work of Postel (2011).

Research questions

Against this background, we can introduce the thesis' research questions, listed in Table 1.2. We focus on the aspects that fit the overall narrative. As our overarching goal is to advance TCPR, we start with a basic question: *why is TCPR so important*?

Many researchers agree that the relevance of TCPR lies in a 'shortcoming' of randomized controlled trials (Elliott, 2010). Effect studies –like these RCTs– are considered to be the golden standard of scientific research. These studies can establish *that* an effect occurred, but as this is an average group-level effect, it is unable to show which aspects of the intervention are related to the change that the intervention establishes. As a consequence, how change occurs as a result of therapy remains a black box, and the ***W**hat **W**orks **W**hen for **W**hom* question (WWWW) cannot be addressed.

Table 1.2: Research questions.

#	*Part*	*Ch.*	*Research question*
i	I	2	Which qualitative methods are used for TCPR?
ii	I	2	Which of these methods have potential for automation?
iii	I	3	Which text mining methods are used for TCPR?
iv	I	4	What differences are there between research disciplines with respect to *automation*?
v	I	4	What differences are there between research disciplines with respect to *explication*?
vi	II	5	How can current state-of-the-art machine learning models be used study e-mail data?
vii	II	6	What makes multilevel models particularly suitable for studying e-mail data?
viii	III	7	How do negative clustering effects affect statistical modelling?
ix	III	8	How can negative clustering effects help to understand *What Works When for Whom* (WWWW)?
x	III	8	Why is *Bayesian Covariance Structure Modelling* a valid approach for studying WWWW?

PART I: TCPR methods

TCPR recognizes that effect studies alone are not the way forward. The history of psychotherapy research is marked by a gradual increase in the understanding of psychotherapeutic change processes (Braakmann, 2015; Orlinsky et al., 2004). Aside from historical relevance, the main research question of TCPR also closely aligns with the clinical practice (Norcross & Wampold, 2011): clinicians are specifically interested in *what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of *circumstances* (Paul, 1967; Tasca et al., 2015, p. 111).

Because almost all treatments rely on the conversation between client and counsellor (Garfield, 2006), we will argue that assessing the *language use* in this

exchange provides an important avenue into the WWWW question. Because *text* is a 'data-format' of language that is straightforward to analyse, it is no surprise that TCPR has a long-standing tradition in the analysis of *transcribed* language (see for example the work of Gottschalk, 1995; Gottschalk & Gleser, 1979).

As our approach to TCPR adheres to the study of natural language, the majority of the relevant TCPR methods is of *qualitative* nature (Elliott, 2010, 2012; Street et al., 2009). To obtain a complete and thorough overview of all available methods, our first research question is *which qualitative methods are used for TCPR* (see Table 1.2).[3] This question will be addressed in chapter 2. As automated methods are becoming increasingly popular, chapter 2 also addresses the question *which of the qualitative methods have potential for automation*. As there also are many *automated* (i.e. 'text mining') approaches for studying texts, chapter 3 addresses the question *which text mining methods are used for TCPR*?

A *systematic review* of the literature is perhaps the best way to answer these questions, because it involves a systematic evaluation and integration of all the relevant literature. This is especially relevant for TCPR, as many researchers formulated their own approach to TCPR, and the literature sprawled in many directions. With these three research questions, part I of the thesis (see Table 1.2) provides an overview of the state-of-the-art TCPR methods. As there are many methods, we use the remainder of part I to reflect on the differences between applied domains such as psychology (chapter 2) and more technically orientated fields (such as computer science; chapter 3). To do so, we identify two trade-offs in chapter 4 that differentiate between technical- and applied-fields: the orientation of *explication*, and the method of *automation*. These two trade-offs form the *automation-explication* framework, which can help to understand *what differences are there between research disciplines with respect to automation and explication*?

PART II: TCPR models

Knowing specifically which automated TCPR methods are available is relevant because technology is on the rise in psychology. Web-based e-mail interventions improved access to psychotherapy for a wider audience (Hoogendoorn et al., 2017), come at low-cost (Schweitzer & Synowiec, 2012), have no to (relatively) short waiting lists (Amichai-Hamburger et al., 2014), are available during a pandemic (Peng et al., 2020), and can be as effective as face-to-face therapy (Andersson & Cuijpers, 2009; Barak et al., 2008; Gainsbury & Blaszczynski, 2011; Howes et al., 2012).

[3] In the remainder of this section, the research questions are all in *italics*, see Table 1.2 for an overview.

The e-mail conversations that we address in the thesis come from web-based psychotherapeutic interventions where the client and counsellor are in different locations, the client follows a (semi-)structured program in an online environment, the client is guided by the counsellor through e-mails for x given weeks, and communication between client and therapist is mainly *asynchronous*[4] (Chester & Glass, 2006; Novak & Pahor, 2017), as online counselling mainly relies on the exchange of e-mail (Rochlen et al., 2004).

Based on the literature reviews we conducted in part I, we found a particular recurring theme: many researchers mention that TCPR would benefit from statistical models that can appropriately analyse "*sequentially dependent observations*" in e-mail data (Elliott, 2010). Many researcher stress that TCPR requires a "*refinement of statistical methods* [...] *to fully account for the multilayered complexity of therapeutic processes*" (Knobloch-Fedders et al., 2015). Models with these properties exist, but as there are disciplinary differences, it is possible that many models are unknown beyond their discipline. In part II, we reflect on the models that are available for TCPR.

As different research disciplines have different (TCPR) preferences, it is not surprising that various fields have a different opinion on what is *ideal* in the context of TCPR. Almost all researchers will agree that a model should be suitable for large scale analysis, can detect therapeutic change (in the exchange between client and counsellor), and can address these changes over time. We discuss machine learning models that fit this description in chapter 5, and statistical models in chapter 6.

In chapter 5 we discuss *how current state-of-the-art machine learning models can be used to study e-mail data* from a web-based intervention for the treatment of AUD. The models we use in chapter 5 are typical examples of models aimed at maximising *accuracy*, rather than providing a good *explanation*. These machine learning approaches contrast with the statistical model we discuss in the next chapter. In chapter 6 we discuss why we favour the use of multilevel models for TCPR. We address *how text mining methods in general can be applied to e-mail data*, and *argue why multilevel models in particular should be used for the analysis of e-mail*. As a proof-of-concept, we (re-)analyse the data from Lamers et al. (2015). We also propose a slight reparametrization of multilevel models to adequately assess therapeutic change.

By discussing the relevant TCPR models in chapter 5 and 6, part II of the thesis presents an overview of the models that are available for TCPR. And even

[4] Asynchronous contact is the most common form of communication in web-based-treatment: client and counsellor talk 'in turns' by responding to each others messages, and communication is therefore time-delayed (Gainsbury & Blaszczynski, 2011).

though we prefer the use of multilevel models for TCPR, there is one shortcoming that underlies all standard multilevel models. Without discussing the content of chapter 6 in detail here, we have to address one specific aspect of the multilevel model here, as it forms the basis of (our discussion of) part III of the thesis.

Multilevel models have the advantage that they can incorporate the *levelled* structure of data. This idea holds potential for TCPR, as change processes are often multifaceted and multi-layered (Knobloch-Fedders et al., 2015). As a counsellor has multiple clients, the clients are said to be 'nested' within a counsellor, as clients with the same counsellor are exposed to a similar treatment environment (Kenny & Hoyt, 2009). By treating counsellors as a level in the hierarchy, multilevel models make it possible to quantify the effect a counsellor has on his or her clients.

PART III: An alternative TCPR model

While trying to model the effectiveness of the counsellors in chapter 6, we found that we were unable to do so because multilevel models[5] have an ill-understood shortcoming. Further explorations of this 'defect' led us deep into almost completely uncharted territory about *negative clustering effects* (i.e. multilevel modelling with 'negative variance components').

It turns out that it is impossible for (standard) multilevel models to assess the negative associations between observations within clusters, which we show in chapter 7. When the data is not independently sampled but positively correlated (i.e. observations are more similar to each other than randomly sampled observations), we use multilevel models. However –as we show in chapter 7– multilevel models can only assess *positively* correlated data, and *negatively* correlated data (i.e. negative clustering effects) are impossible to address with standard multilevel models.

We also use chapter 7 to illustrate that the data we used in chapter 6 contains these negative cluster effects. We present a simulation study to show *how negative clustering affects statistical modelling*, and demonstrate the effects negative clustering has on parameter estimation, the type-I errors, and hypothesis testing.

Chapter 7 leaves one important question open: if multilevel models cannot assess negative clustering effects, how should these effects then be treated? In chapter 8, we give a gentle introduction to the ***B**ayesian **C**ovariance **S**tructure*

[5] In chapter 6 we refer to these models as *multilevel models*, in chapter 7 as *linear mixed models*. *Multilevel models* are also known as *hierarchical linear models*, *mixed models*, *nested data models*, *random coefficient*, and *random-effects models* (Raudenbush & Bryk, 2002a).

***M**odelling* (BCSM) *framework* (Fox et al., 2017; Klotzke & Fox, 2019a, 2019b). We show why *BCSM is a valid approach for studying negative clustering effects*, and provide a gentle introduction to the (mathematical) background of BCSM. We also apply BCSM to the Lamers et al. (2015) data (we used in chapter 6) and show that we were unable to estimate the counsellor effect in chapter 6, because the counsellors had a negative clustering effect on their clients (also discussed in chapter 7). We discuss that this effect is –in fact– an individualized effect, and argue that *negative clustering effects contain information about the WWWW question*.

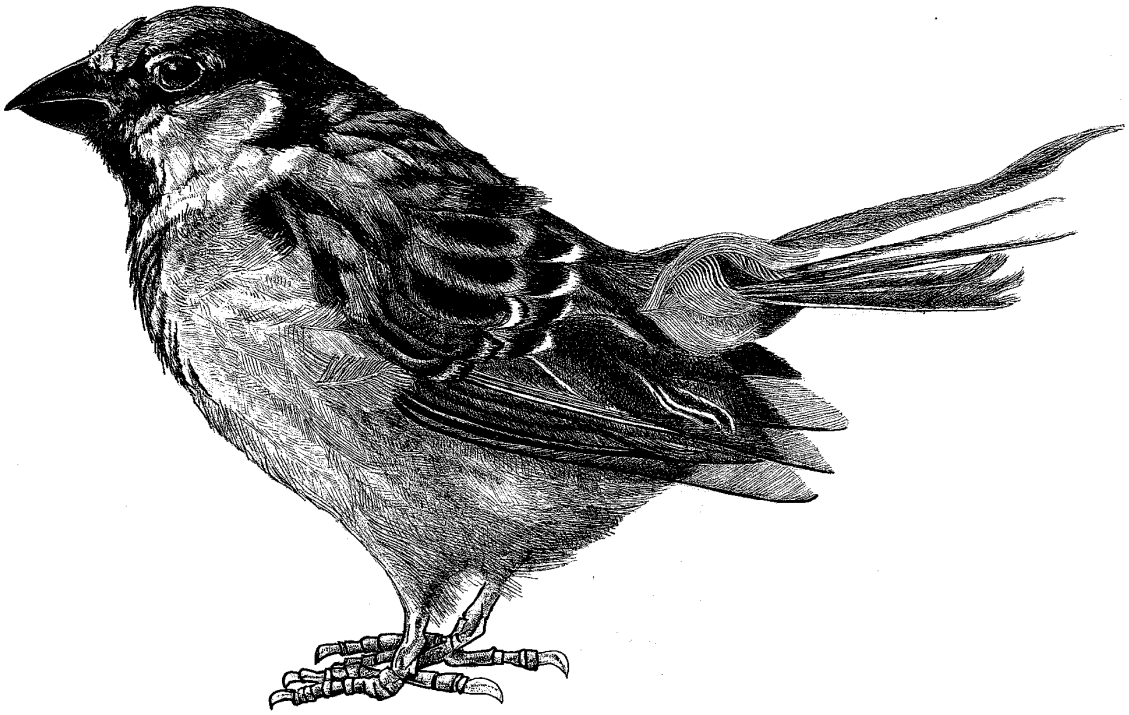
The final chapter: a general discussion

A roadmap through the thesis would be helpful: in part I we introduce TCPR, a research discipline devoted to studying the WWWW question, and present an overview of the available TCPR *methods*. In part II, we discuss different TCPR *models*, and argue for the importance of multilevel models. In part III we introduce BCSM, an *alternative* TCPR *model* capable of assessing negative clustering effects, which are intricately related to the WWWW question. So, in part I we synthesize what is already done, in part II we apply existing models to e-mail counselling, and in part III we explore a new approach.

We conclude the thesis with chapter 9, where we summarize (and where possible) integrate part I, II, and III. We also addresses the general limitations of our research projects, and provide an overview of the implications for future research. We summarize the results of the thesis by posing seven research question (see Table 9.1).

Part I

TCPR methods



What
Works
When
for
Whom

The impetus for this article is to make the case for more research on mechanisms of therapeutic change. **This is not a new plea**.

Kazdin and Nock (2003)

What
Works
When
for
Whom

2 Towards Text Mining Therapeutic Change: A Systematic Review of Text-Based Methods for Therapeutic Change Process Research

Abstract

Therapeutic Change Process Research (TCPR) connects within-therapeutic change processes to outcomes. The labour intensity of qualitative methods limit their use to small scale studies. Automated text-analyses (e.g. text mining) provide means for analysing large scale text patterns. We aimed to provide an overview of the frequently used qualitative text-based TCPR methods and assess the extent to which these methods are reliable and valid, and have potential for automation. We systematically reviewed *PsycINFO*, *Scopus*, and *Web of Science* to identify articles concerning change processes and text or language. We evaluated the reliability and validity based on replicability, the availability of code books, training data and inter-rater reliability, and evaluated the potential for automation based on the example- and rule-based approach. From 318 articles we identified four often used methods: the *Innovative Moments Coding Scheme*, the *Narrative Process Coding Scheme*, the *Assimilation of Problematic Experiences Scale*, and *Conversation Analysis*. The reliability and validity of the first three is sufficient to hold promise for automation. While some text features (content, grammar) lend themselves for automation through a rule-based approach, it should be possible to automate higher order constructs (e.g. schemas) when sufficient annotated data for an example-based approach are available.

Keywords: Therapeutic Change Process Research (TCPR), systematic review, text-based qualitative methods, text mining, automation

Smink, W. A. C., Sools, A. M., Van der Zwaan, J. M., Wiegersma, S., Veldkamp, B. P., & Westerhof, G. J., (2019). Towards Text Mining Therapeutic Change: A Systematic Review of Text-Based Methods for Therapeutic Change Process Research. *PLoS One*, 14*(12)*, e0225703.https://doi.org/10.1371/journal.pone.0225703

Introduction

Big data and automated analysis methods are nowadays so omnipresent that they change traditional scientific research practices. Publications investigating new possibilities that these methods bring forth appear almost on a daily basis in every scientific discipline, and the field of psychotherapy research is no exception (Owen, 2013). The rising popularity of machine learning methods to automatically analyse large bodies of data or texts accelerates research not only by allowing new kinds of research questions to be answered, but also by re-establishing the relevance of known questions which require the analysis of (text) data.

As early as Freud's talking cure, the importance of looking at language to understanding the therapeutic process has been recognized (Smink, Fox, et al., 2019).[1] The idea that the verbal exchange between counsellor and client contains important ingredients of therapy fuelled psychotherapy research (Imel et al., 2015). There is a long-standing tradition of studying the linguistic 'products' of therapy (e.g. homework exercises, diaries, transcripts) in order to understand therapeutic change. The underlying idea is that the assessment of natural language use can reveal the process and changes over the course of therapy (He, 2013). Thus, counsellor and client transcripts could potentially be a direct observation of the therapy process (Elliott, 2012; Gelo et al., 2012; Murphy et al., 2015). From this perspective, the psychotherapeutic process is considered a highly structured form of interaction, of which many important aspects are of linguistic nature.

Text-based therapy research has mainly relied on manual coding and human interpretation (Elliott, 2010). With the rise in available therapeutic texts in this digital age, automated analysis is making its entry into therapy research. There is now a growing number of studies on automated screening and diagnosis (Adler, 2012; Andersson & Cuijpers, 2009; Atkins et al., 2014; Tanana et al., 2015), and interest in automated analysis of the therapy change process is also picking up (cf. Howes et al., 2014). In our view, automated analyses in this field did not yet reach full potential because there are privacy and ethical concerns with sharing data from clients and patients (Bennett et al., 2010; L. Bishop, 2009), and because the field is pragmatically organized. This means that data driven approaches prevail (there are –of course– exceptions, cf. Cariola, 2015; Mergenthaler, 1996; Murphy et al., 2015), and that –for some methods– the availability of data for automated analysis determined the research questions and approaches, rather than that these decisions are based on psychological theory and research.

[1] Smink, Fox, et al. (2019) is chapter 6 of this thesis.

We propose that human interpretation and computer-based automated analysis can benefit from each other, and each have their distinct function. The large body of existing theories, models and methods for text-based analyses developed for understanding therapeutic change are currently underutilized. Yet, we would argue that these theories, models and methods are crucial for generating meaningful questions for automated analysis, and for a meaningful interpretation of patterns detected by a computer. Vice versa, the idea is that computers can be trained to perform (at least part of) the very labour intensive work of coding large bodies of text. This would enable the testing of hypotheses at an unprecedented scale, which is difficult to do with many of the existing methods that assess therapeutic change processes (Elliott, 2010).

Therefore, we have the ambition to align automated analyses with existing text-based methods for therapeutic change processes. A prerequisite of a well-founded, meaningful development of automated text analysis is an overview of the available qualitative methods developed for the purpose of understanding psychotherapeutic change. Towards that end, we present our systematic review of literature on relevant peer-reviewed, published text-based methods for studying therapy change. In the remainder of this introduction, we first describe the field of *Therapy Change Process Research* (TCPR; Smink, Fox, et al., 2019), followed by a description of what *text mining* is, and what text mining has to offer in the context of understanding therapeutic change processes. We conclude with a discussion of *rule-* and *example*-based approaches for text mining.

Therapeutic Change Process Research

Over a third of the people in most countries report problems at some time in their life which meet criteria for diagnosis of one or more of the common types of mental disorder (WHO; Degenhardt et al., 2018). For example in the Netherlands, mental health problems contribute to about a quarter of the losses in Dutch health-life years (de Graaf et al., 2011). In light of these statistics, it is not surprising that more than a thousand different psychotherapies have been developed (Garfield, 2006). Hundreds of studies already demonstrated that professional treatment can help people change in desired ways (Lambert & Bergin, 1994). To ensure that these therapies are supported by sufficient empirical evidence the APA adopted a resolution on the effectiveness of psychotherapy (L. F. Campbell et al., 2013).

However, progress in psychotherapy research is not made by only demonstrating the effectiveness of a treatment. In spite of thousands of studies publishing the outcomes and effects of therapies (Barkham et al., 1993; Elliott, 2012;

Nock, 2003), the most intriguing questions remain: why and how do treatments work for whom? Studies aimed at average effects at group level fail to understand vast individual differences in responsiveness to therapy (Kazdin & Nock, 2003; Kent & Hayward, 2007; Norcross & Wampold, 2011; Tasca et al., 2015). Therefore, TCPR "*aims to identify the mechanisms through which psychological treatments bring about positive and therapeutic change*" (Smink, Fox, et al., 2019).

TCPR is accordingly defined as "*the scientific investigation of what occurs during psychotherapy, with regard to its clinical meaningfulness; in other words, it investigates the process through which clinically relevant changes occur within psychotherapy*" (Gelo & Manzo, 2015, p. 259). Smink, Fox, et al. (2019) noted that various names and definitions are used throughout the literature: *Change Process Research* (CPR; Elliott, 2010; Greenberg, 2007), *Psychotherapy Process Research* (PPR; Gelo et al., 2012), and some of the early works simply refer to '*change*' (cf. Braakmann, 2015; Hill & Corbett, 1993; Shapiro, 1995). To emphasize that we are dealing with change resulting from therapy, we propose to describe change processes as *Therapeutic Change Process Research* (TCPR; Smink, Fox, et al., 2019). We use the terms *therapy* and *therapeutic* synonymously with *psychotherapy* and *psychotherapeutic*. 'Change' in TCPR then refers to the (positive) improvement in the client that is the result of psychotherapy (i.e. psychotherapeutic change). Although it is conceivable that therapeutic interventions (also) have negative effects, we limit ourselves here to the positive and beneficial effects of therapy.

Greenberg –who formally defined [T]CPR in 1986– was, together with Carl Rogers (1961), among the firsts to argue for the importance of understanding change. Since then, many different TCPR methods have been developed (Braakmann, 2015; Elliott, 2010; Wallerstein, 2001). Like other psychological research methods, TCPR methods also vary in their reliance on forms of statistical inference (Mörtl & Gelo, 2015). In a rather broad definition, qualitative psychological methods mainly rely on the interpretation of natural language (Hill & Corbett, 1993). Contrasting are quantitative linguistic TCPR research methods, that in practice usually equate to forms of counting of words (cf. *Linguistic Inquiry and Word Count*, LIWC; Pennebaker, Boyd, et al., 2015). LIWC, pronounced as the English name *Luke*, appears to be one of the forefront of the quantitative methods; in our current work, we however focused on the qualitative approach. For a more complete overview of (the differences between) quantitative and qualitative methods, see Gelo and Manzo (2015, p. 259).

Most examples of qualitative approaches adhere to the interpretative study of the natural language used in therapeutic interaction (Elliott, 2010, 2012; Street et al., 2009), and are based on the assumption that word use reflects various psychological processes and change mechanisms (Arntz et al., 2012). For example,

Wynn and Wynn (2006) identified cognitive, affective, and sharing empathy in sequences of therapeutic talk.

Over time, qualitative and quantitative approaches to TCPR developed into rather independent and different communities of researchers (Braakmann, 2015; Salvatore et al., 2012; Wallerstein, 2001). By systematically reviewing the qualitative TCPR approaches, we intend to present the state-of-the art, allowing for more integration of the two approaches. Clearly, there is room for doing so: the recent increase in web-based interventions (there is a variety of different names for *online therapy methods*, see Barak et al., 2008; Oh et al., 2005), like e-mail supported life-review interventions (e.g. Amichai-Hamburger et al., 2014; Lamers et al., 2015), generate textual data directly, discarding the need for transcriptions (Chung & Pennebaker, 2007; Imel et al., 2015), omitting this labour-intensive process. Also, data of therapeutic sessions are nowadays more easily collected than ever (Andersson & Cuijpers, 2009; Andrews et al., 2010; Hoogendoorn et al., 2017).

Nevertheless, the increased availability of these data did not lead to a substantial increase or popularization of TCPR research in general (Gelo & Manzo, 2015): all developments resulted in larger availability of data, although this does not also automatically result in larger *access* to datasets for research. Partly, this is because the privacy of respondents is protected by ethical protocols and strict legislation, which prohibits data sharing and making datasets publicly available for TCPR (Bennett et al., 2010; L. Bishop, 2009). Another reason –and one that we shall discuss in detail– is because "*the technology for evaluating psychotherapy* [for the qualitative field] *has remained largely unchanged since Carl Rogers first published verbatim transcripts in the 1940s*: *sessions are recorded and then evaluated by human raters*" (Atkins et al., 2014). Indeed, development of the automated research methods is –relatively– slow (in comparison to other fields in Psychology and Psychiatry; Abbe et al., 2016).

Text mining therapeutic change

As some argue that the amount of textual data currently available makes human evaluation no longer a feasible, valid or reliable method given realistic time- and budget constraints (Basit, 2003; Imel et al., 2015; Snow et al., 2008), it should not come as a surprise that text mining methods appear to be on the rise in psychology (Sools et al., 2019).[2] Text mining refers to a general methodological framework that includes several automated methods to analyse large corpora of texts. Practically, text mining approaches in psychology include counting

[2]Sools et al. (2019) is chapter 3 of this thesis.

words, identifying topics, and coupling the terms to a domain-specific ontology (Hoogendoorn et al., 2017). As text mining combines techniques and methods from many disciplines –including linguistics, statistics, computer science, natural language processing (NLP), artificial intelligence, information retrieval and data mining– it is not surprising that terms referring to the automatic extraction of information from text are used interchangeably, such as *text mining* and *NLP* (Jurafsky & Martin, 2014). Therefore, text mining is broad umbrella term that refers to a general methodological framework that includes several automated methods to analyse large corpora of texts.

We recommend novel and aspiring practitioners of text mining the works of R. Feldman and Sanger (2007) and Jurafsky and Martin (2017), and Manning and Schütze (1999). We recommend aspiring text mining practitioners the NLTK library, which is available in the programming language Python, and has an extensive step-by-step manual written by Bird et al. (2009, which can also be used by those with little to no familiarity to programming or Python). We recommend Sools et al. (2019) to those especially interested in text mining TCPR.

It is possible to identify a framework of studies that model change processes similar to what we aim to achieve by combining text mining and TCPR. For our purpose, we distinguished these works as *theory*- and *data*-driven approaches. The most well-known automated theory-driven text analysis tools is perhaps LIWC (Pennebaker, Boyd, et al., 2015). This text analysis program counts words in psychologically meaningful categories, and because it relies on previous research and theory to establish the relevance of the word categories it is considered a theory-driven method. Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, such as showing attentional focus, emotionality, social relationships, thinking styles, and individual differences (Pennebaker et al., 2003).

Data-driven techniques are often developed to be broadly applicable, and regularly apply standard text mining tools to data with less reliance on a specific text analysis theory developed for that field. An example of a data-driven method is topic modelling, which refers to the use of type of statistical to discover abstract topics occurring in a collection of documents (Blei et al., 2003). For example, Atkins et al. (2012) used this technique to analyse transcripts of therapy sessions from couples in a randomized trial, where the topic model establishes which words tend to occur together in transcript documents (e.g. *mom*, *mother*, *dad*, *sister*, and *brother* all belong the topic *family*).

The distinction between theory- and data-driven methods is then characterized by the extent to which methods incorporate theoretical knowledge. LIWC is a method that relies mainly on theory, whereas topic models are mainly data-

driven. The characterization of theory- or data-driven methods becomes relevant in the context of what we would call the distinction between the *rule*- and *example*-based approach.

The former is based on annotated data and coding schemes, whereas the latter is based on linguistic information and text feature extraction. For example, Pfäfflin et al. (2005) examined patterns across therapy by labelling utterances and sessions with several client-counsellor relationship variables. These manually labelled texts were then used for the text mining analyses. The rule-based approach is characterized by a more or less 'automatic' extraction of text features from texts through a set of pre-defined rules. For example, Anderson et al. (1999) determined several process verbs, and counted their occurrence in therapeutic sessions. Another example is Atkins et al. (2014), who rely on topic-modelling to cluster sessions based on similarities in word use in the in-therapeutic utterances.

Within the theory-driven approaches, we distinguish between those relying on rule- and example-based approaches. As we will argue, this differentiation within the text mining field is relevant to the question how TCPR research can (best) be automated. The distinction between theory- and data-driven methods is not made formally: the majority of methods is a hybrid, and –if one intends to classify methods based as theory- or data-driven– it is perhaps best to place methods a continuum where theory- and data-driven mark the edges.

Rule- and example-based approaches to automation

There are multiple approaches for text mining; we will discuss and highlight the importance of rule- and example-based approaches. Especially rule-based models are best understood in their historical context, but we keep the discussion of the history of the field to a minimum here (we refer the interested reader to Jurafsky & Martin, 2014).

The rule-based approach

The earliest applications of what is now known as text mining come from computer scientists who –just after the Second World War– tried to model, analyse and understand speech and written natural language through rule-based language models (Johnson, 2009; Jurafsky & Martin, 2014). The work of these pioneers emphasized the core of rule-based models: some *input* (a text or verbatim) is mapped to an *output* (a label or a category) through some *function*. Rule-based language models describe a set of models that explicitly define the relation between input and output through a set of hand-coded rules for the

function (Mykowiecka et al., 2009). The rule-based approach thus mandates that the researcher explicitly specifies the routine by which lexical clues will be obtained, or that the researchers specifies exactly in advance which words contain relevant information.

For example, a comprehensive search string (a '*regular expression*'; Brzozowski, 1964; McNaughton & Yamada, 1960) was used to detect whether an utterance contains a check question, suicide ideation, appreciation or surprise (Althoff et al., 2016). Similarly, decision trees with hand-crafted rules were used to classify sentences to open-ended questions (Gallo et al., 2015). The rule-based approach was also used to distinguish differences between linguistic measures and outcome measures was examined in high and low verbalized affect segments (Anderson et al., 1999). Others used the rule-based approach to show the correlation between verb repetition and differences in affective arousal (Halfon et al., 2017). This approach comes with the advantage that the researcher has direct control on what is extracted from the text. The other advantage is that theoretical knowledge can be directly applied: researchers often have a good idea on which words or expressions are related to their outcome of interest. The disadvantage is that when a researcher does not have theory to dictate what is important, it can be difficult to decide which words or information is are 'more' relevant than others.

Another disadvantage that limited the practical use of rule-based models is that the number of rules necessary to model natural language needs to be extremely large. Over the years, scientists from different fields (such as computer science and electrical engineering; Jurafsky & Martin, 2014) began experimenting with language models that were not based on comprehensive sets of rules, but that 'learned' to model language based on 'raw' examples from texts. Around 1990, this led to what many refer to now as a 'statistical revolution' (Martinez & Martinez, 2015); example-based (machine learning) models became more prominently featured in text mining than rule-based models (Johnson, 2009; Manning & Schütze, 1999).

The example-based approach

Around the 90s, computational resources and the availability of data both greatly increased (for example the large *Linguistics Data Consortium* became available; Jurafsky & Martin, 2014; Liberman, 2002), making way for example-based models, which typically demand more data and computational power than rule-based models. It turned out that the probabilistic data-driven models from statistics and machine learning were better suited for modelling natural language (Sofaer et al., 2019). In about the span of a decade, example-based models completely

took over the field (Martinez & Martinez, 2015).

To sharpen the contrast with rule-based models, we propose to call these models example-based, instead of 'statistical' or 'machine learning' models. The core of example-based models is that they rely on statistical inference to automatically learn the 'rules' of a language through the analysis of large corpora of typical real-world examples (instead of through specific hand-written rules; M. Bates, 1995). More formally: the function is 'learned' by providing an exampled-based algorithm with specific examples of how the input and output should be associated.

The example-based approach is characterized by the application of text mining algorithms in order to find meaningful relations between human annotator derived labels (or ratings) and lexical cues in the data. The example-based approach mandates sufficiently large hand-coded datasets where differences in the text are related to differences in the outcome data (Basit, 2003). For example, language models trained on (i.e. 'machine learning') hand-labelled counsellor utterances for low and high empathy sessions are used to predict empathy in sessions (cf. Xiao et al., 2016). Annotated data are also used to automatically distinguish 'change' and 'sustain talk' in the client and counsellor utterances in motivational interviewing (Tanana et al., 2015). The practice is clear: without specification of any formal rule (that characterizes the rule-based approach), the example-based approach is able to learn, classify and predict labels with satisfactory accuracy if sufficient hand-coded data is available.

This approach comes with two drawbacks for the psychological practice. First, it requires a lot of hand coded data, which is not always available (because data sharing is not always allowed under strict privacy regulation Bennett et al., 2010; L. Bishop, 2009). Second, the construction of such datasets is extremely expensive in both annotator-hours and cost (Snow et al., 2008). Since the performance of many natural language processing tasks is limited by the amount and quality of data available to them (Banko & Brill, 2001), one promising alternative for some tasks is the rule-based approach.

Note that our distinction between example- and rule-based approaches does not mean that these two approaches are mutually exclusive. Figure 2.1 reflects our view on the matter: the two approaches form the ends on a spectrum. A method can rely on both approaches for automation, but usually one of the two can be preferred over the other when a first attempt is made at automation.

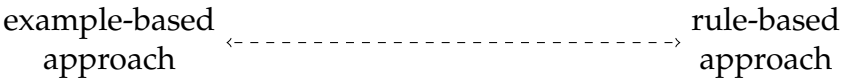


Figure 2.1: Research methods in TPCR can automated based on the extent by which they rely on a rule- or an example-based approach for automation.

Research goals

TCPR aims to connect in-therapy change processes to outcomes. Qualitative instruments are commonly used to study the linguistic products of therapy. However, due to the dependence on human interpretation these methods are limited in analysing the large bodies of text that are nowadays available, limiting their use to small scale research. We therefore advocate the combination of TCPR and text mining. Towards that end, we present a systematic review in which we aim to provide an overview of the commonly used methods, peer-reviewed qualitative text-based TCPR methods, assess to what extent these methods reliable and valid, and assess the extent to which these methods are automatable based on the rule- or example-based approach.

Method

A commonality of TCPR is the frequent co-occurrence of 'process research' and 'change process'. We expressed interest in psychological treatments through the queries 'psychotherapy', 'counselling', and 'treatment'. We identified qualitative TCPR through the queries 'language', 'text' and 'transcripts', including 'narrative', 'discourse' or 'conversation' analysis, see Figure 2.2 for an overview of our search query.

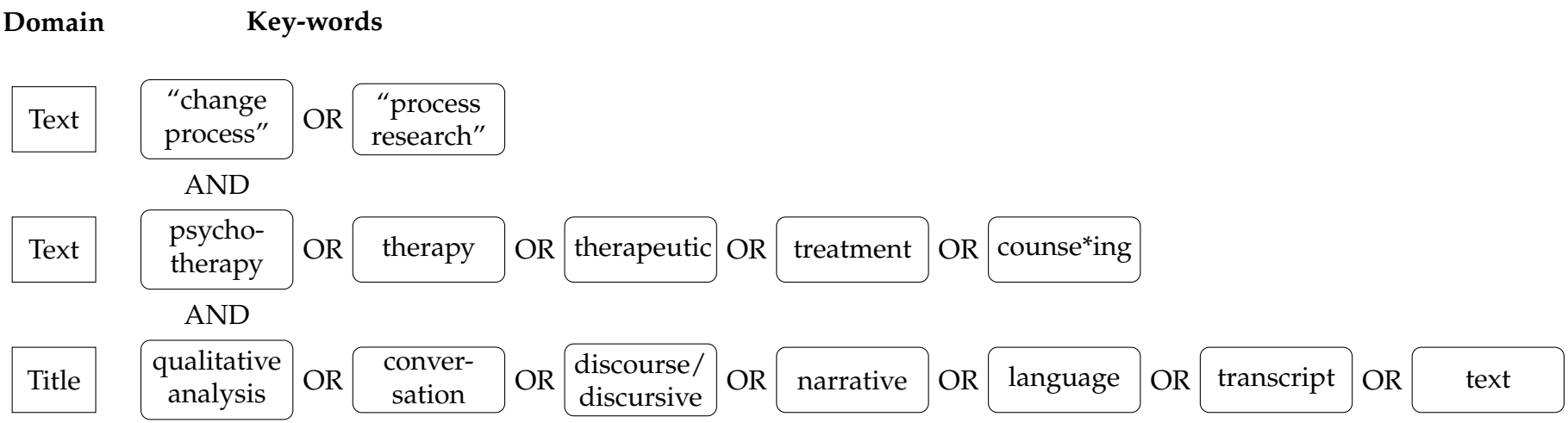


Figure 2.2: Search query used to search the PsycINFO, Web of Science and Scopus databases. The key-words have blocks with round corners; blocks with sharp corners indicate the search query's domain (assessment of the title, or the full-text).
$*$ indicates the use of a wild card, when different forms of spelling can be used.
/ indicates that two words are treated as equivalents.

We used the first and third block of Figure 2.2 to search titles, abstracts and full-texts, the second block was used to search only titles. As treatment-related queries frequent in all psychotherapy articles, inclusion of these terms for full-text searches led to an increase of many articles not directly related to the research question. To ensure good coverage, we included several important and impactful TCPR publications, for which we consulted TCPR experts.

Databases

We searched three scientific databases: *PsycINFO*, *Scopus*, and *Web of Science*. PsycINFO should contain many TCPR records as it centres on *psychology*, the *behavioural* and the *social sciences*. We also used Scopus, as it contains the *MEDLINE* databases, which also span *psychiatry* and *medical psychology*. To also include the *humanities*, we also searched Web of Science.

Inclusion The inclusion criterion is that a study has to report on a *TCPR instrument* through the *assessment of language* or *text-components*, such as transcripts, diaries, e-mails, psychotherapeutic assignments.

Exclusion Reasons for exclusion besides not meeting the inclusion criterion were: not a scientific publication (e.g. commissioned report, organizational project paper, book or book review); not an empirical study (e.g. theoretical perspective on change or therapy); aimed at another change process than therapy (e.g. career counselling, flourishing); not a target group with common mental health disorders (e.g. stuttering, sexual offenders HIV patients); and not measuring (individual) client-counsellor interactions (e.g. group-therapy, family-therapy).

Identification and selection of methods and studies

After removing duplicate articles, the first and last author independently screened all titles and abstracts for inclusion and exclusion criteria. Identifying the articles to exclude turned out to be relatively straightforward. Agreement upon inclusion was not so easily reached, and we calculated Cohen's $\kappa < .70$. This statistic mainly reflects that TCPR-related literature was addressed by multiple different disciplines under a variety of different names, making it difficult to reach agreement upon inclusion.

One of the two screeners was an experienced TCPR-researcher, the selection of this screener turned out to mark more articles for inclusion. To avoid the risk of excluding relevant articles –which is the largest risk when κ is below

the cut-off point of .70– we decided to include all articles that either one of raters selected for inclusion. All literature marked for inclusion was then fully read. Both screeners labelled the articles with the method that the authors used. Finding the frequently used methods then –essentially– boiled down to counting all the methods that were found.

We chose to only give an elaborate description the methods that were mentioned more than twice in the literature. We made the assumption that the methods that were used only once or twice could not have had an lasting impact on the TCPR field.

Data analysis

We assessed the full-texts of the articles that used frequently occurring methods in three steps, one for each of the three research questions. Data analysis in these three steps was conducted by the second author and checked by the first author. As basis for the analysis we identified one key article for each of the methods. The first article where the method was described in detail, or the article that was referred to by all other articles using that method, the 'source'-article. This was supplemented by an analysis of articles citing the key article and/or using the same method. We choose the key article to be the article that first proposed the method, or contained the most information on how to specifically apply the method.

Step 1. Description of the methods Here we describe, mainly on the basis of the key article, if and how the theoretical background and main concepts of the method are provided by the authors of the method. We paid attention to how explicit and elaborated underlying theoretical models and concepts were described.

Step 2. Assessment of quality criteria We looked at the reliability and validity of the included methods. For our assessment, we first analysed if and how authors provided argumentation to explicitly address the validity and reliability of their method. This analysis of explicit accountability for the quality of the methods was complemented by our own analysis of more implicit evidence for the quality of the methods either within the key article, or by reference to other articles adopting the same method. For assessment of the validity of a method, we looked at *internal* and *external validity*.

Validity We deemed a method *internally valid* to the extent that claims and constructs were substantiated with existing theories and models, and/or empir-

ically validated using transcripts and examples. Internal validity increases to the extent that an underlying theoretical framework or model (for the method as a whole and/or for key constructs to be measured), is made explicit and detailed by authors. In anticipation of the question about the automatibility of the method, we additionally described whether applicability included the availability of linguistic markers for the identification of labels (this also added to the reliability of the measure).

The *external validity* of a method increased when transferral to other contexts, client groups, or therapies is made plausible. As indications for transferability of the method, we looked at explicit argumentation by the authors, and for evidence that the method has been used in various applications. In addition, we looked for more implicit indications for transferability, such as the provision of points of comparison which enable analogical reasoning necessary to discern commonalities and differences with other cases where the method could have been applied (Smaling, 2003).

Reliability We deemed a method reliable when the description of the method demonstrated *consistency* (the extent to which data can be analysed independent from other raters and arrive at the same conclusions) and *transparency* (the possibility to virtually replicate the procedures, failures, and successes of the original study). We assessed the consistency of a method based on the reporting of the inter-rater reliability score of the coding scheme (if provided), and the transparency of the method depending on the presence of a manual or coding system with good labels, examples of texts and a clear operationalization.

Step 3. Assessment of automatibility Because these qualitative methods were not originally meant for automation, we deduced the potential for automation from the combination of traditional criteria for methodological quality, e.g. reliability and validity. In our view, reliability is a necessary condition for a method to have potential for automation: if human raters cannot reach good reliability, automated methods cannot be expected to do better. While this may generally apply to all forms of text mining, we made a distinction between a rule- and example-based approach for text mining, to let the qualitative research practice better align with the nature of text mining methods.

An example-based approach to text mining requires the availability of a good coding scheme with high inter-rater reliability. Based on large amounts of manually coded data, a computer can be trained to repeat the analysis. The accuracy of the computer in analysing which text segments are associated with which codes, can then be tested using a test set (again consisting of annotated data).

Table 2.1: Number of articles that mention different TCPR methods.

	Methods	*Articles*
Once	50	50
Twice	8	16
Often used	4	29
Total	62	95

Note. The 62 methods were disaggregated to whether they were mentioned once, twice, or more than twice in the literature. We assessed the full-texts of 45 articles (out of 95 in total, 47.4%) of methods that were mentioned twice or more often.

A rule-based approach to text mining on the other hand, does not require any manual coding, but rather depends on the availability of linguistic markers for TCPR-related constructs. The more information about word use, grammar, or other linguistic features form text are provided, the higher chances that a suitable text mining tool can be identified (or developed) for mining the construct.

Results

Our search resulted in 192 articles in Scopus, 167 in PsycINFO, and 100 in Web of Science, see Figure 2.3. After removing the 194 duplicates, the first and last authors independently screened (the methods described by) 318 articles. Independently, both raters selected (in total) 95 unique articles. These 95 articles described a total of 62 methods that met the inclusion criteria in the opinion of either one or both of the authors, see Table 2.1. 80.6% of these methods were only mentioned once, covering 52.6% of all the included literature (percentages can be calculated from Table 2.1). The other 12 methods, which are described by 45 articles see Table 2.2 (and Figure 2.3), therefore also covered (slightly less than) half of the included literature, but are far more likely to have impacted the field. Eight of these methods were used only twice (see Table 2.1 and 2.2); the other four methods occurred more than twice (see Table 2.2), and cover 64.4% of all methods that occur more than once (see Table 2.1). After reading the full-texts of the 29 articles describing the often used methods (see Figure 2.3), we entered $N = 7$ articles in our study (see Figure 2.3), describe $N = 4$ TCPR methods (see Figure 2.3).

Terminology

We included these four frequently used methods (by including their –in total– 7 manuals) in our review: *Assimilation of Problematic Experiences* (APES; Stiles et al., 1990; Stiles et al., 1991), *Innovative Moments Coding Scheme* (IMCS; Gonçalves et

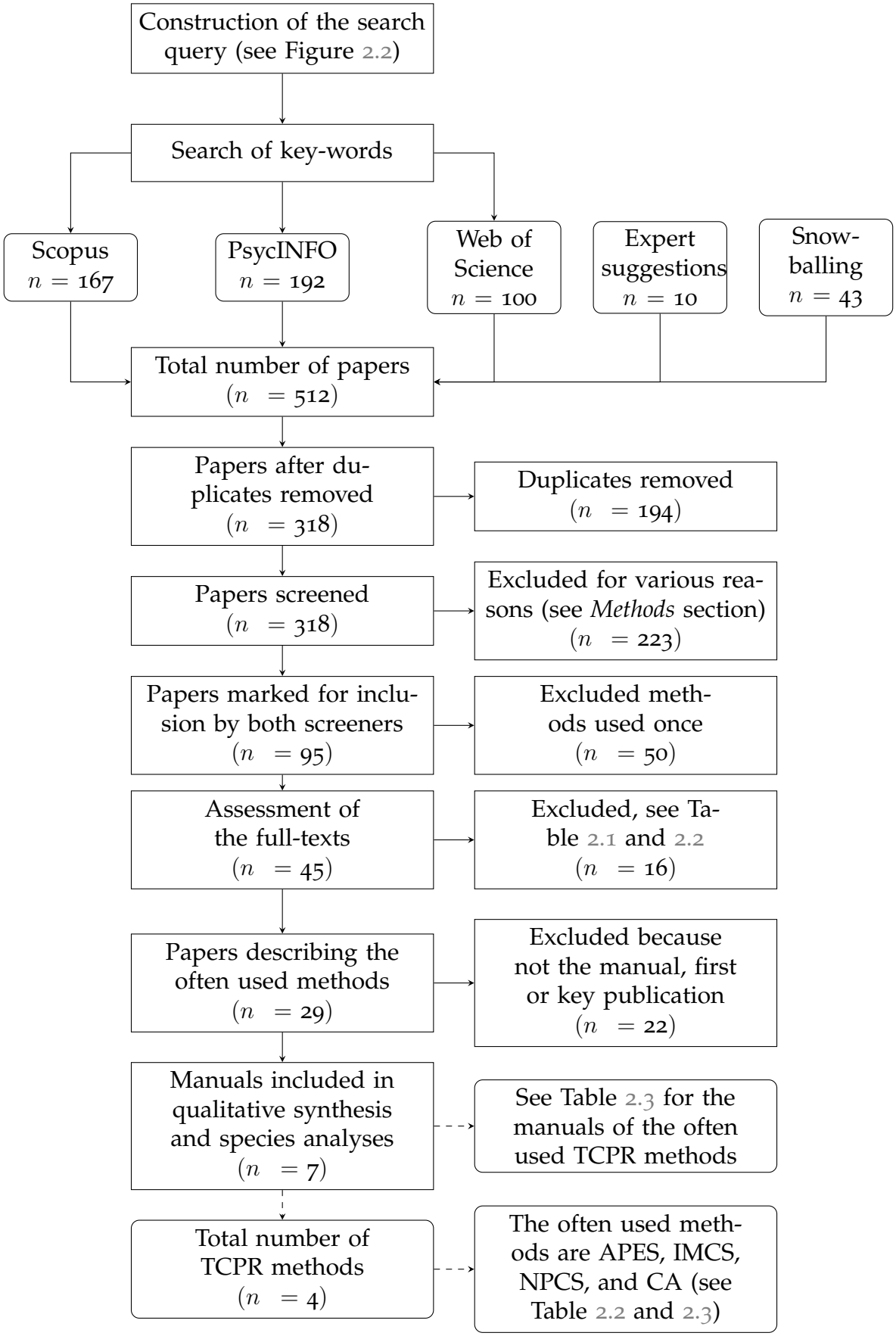


Figure 2.3: Flowchart of the information through different phases of the systematic review. In total, we included 7 articles describing 4 methods, see Table 2.3 for the abbreviations of the methods and the corresponding articles.

Table 2.2: Methods and how often they were encountered in the literature search.

#	*Methods*	*Abbr.*	*Count*
1	Innovative Moments	IMCS	16
2	Conversation Analysis	CA	5
3	Assimilation Analysis	APES	4
4	Narrative Process Coding Scheme	NPCS	4
5	Comprehensive Process Analysis		2
6	Core Conflictual Relationship Theme		2
7	Discourse Analysis		2
8	Metaphor Analysis		2
9	Return-to-the-problem markers		2
10	Structural Analysis of Social Behaviour		2
11	Thematic Analysis		2
12	Therapeutic Collaboration Coding Scheme		2

Note. The first four methods are the methods that were used most often. We included these four methods in our review. In total, we found 29 articles describing the four often used methods, see Figure 2.3. We also included the abbreviations of their manuals and codebooks, see Table 2.3.

al., 2010; Gonçalves et al., 2011), *Narrative Process Coding Scheme* (NPCS; Angus et al., 1996), and *Conversation Analysis* (CA; Peräkylä, 2012; Voutilainen et al., 2011), with IMCS clearly outranking the other three methods in terms of frequency (16 articles using this method as opposed to 5 times conversation analysis, and 4 times the other two methods, see Table 2.2).

We extensively studied 7 articles, describing 4 methods, see Table 2.3. The terminology that we used to denoted the methods and articles will be in similar fashion, to which we (sometimes) refer interchangeably. To avoid confusion, we explicitly described the methods and their manuals in Table 2.3.

Coding schemes

We answer the research questions for each of these four methods. The first (description of background and main concepts) and second research question (assessment of their validity and reliability) culminate in the third question (potential for automation). This assessment was done by the first two authors. See Table 2.4 for an overview our findings on the quality assessment of these methods. We will discuss the content of Table 2.4 in detail in the four following sections, one for each method, starting the most often used methods (see Table 2.2). If training data is present, we checked whether a references was made to this training set (for example Gonçalves et al., 2011, refer to their training data).

Table 2.3: Manuals of the often used methods.

Short	*Scale*	*Manual*
IMCS	Innovative Moments Coding System	Gonçalves et al. (2010) Gonçalves et al. (2011)
CA	Conversation Analysis	Peräkylä (2012) Voutilainen et al. (2011)
APES	Assimilation of Problematic Experiences Scales	Stiles et al. (1990) Stiles et al. (1991)
NPCS	Narrative Process Coding System	Angus et al. (1996)

Note. The seven manuals belong to the four methods mentioned in Table 2.2 and Figure 2.3.

Table 2.4: Overview of the quality of the frequently used methods.

Method	*External validity*	*Internal validity*	*Reliability*	*Annotated data*	*Linguistic*
IMCS	Medium	Medium	High	High	?
CA	Low	Low	Low	Low	Low
APES	High	High	Medium	Partly	Promising
NPCS	?	Medium	High	High	Promising

Note. We abbreviated the four often used methods: *Assimilation of Problematic Experiences* (APES; Stiles et al., 1990; Stiles et al., 1991), *Innovative Moments Coding Scheme* (IMCS; Gonçalves et al., 2010; Gonçalves et al., 2011), *Narrative Process Coding Scheme* (NPCS; Angus et al., 1996), and *Conversation Analysis* (CA; Peräkylä, 2012; Voutilainen et al., 2011). "?" refers to *unknown*, or *not assessable in its current form*.

Innovative Moments Coding Scheme

IMCS screens therapy sessions for innovative moments (IMs). IMs are defined as "*episodes in which the person did, thought, imagined, or felt something different, or related to others in a new way, distinct from the rules that the problematic self-narrative 'prescribes' for his or her life*" (Gonçalves et al., 2010, p. 107). An IM is a concept derived from narrative therapy. The underlying narrative approach to change is contrasted with traditional psychotherapeutic models (e.g. social rather than psychological, focused on re-authoring of alternative stories rather than repairing deficit). While starting out as an alternative to traditional therapeutic models, the method is proposed to have wide applicability even extending to the study of change in everyday life. The authors make their aim of generalizing the method to diverse psychotherapeutic models and different populations explicit. A further claim for the external validity of the method is presented by emphasizing its compatibility with mixed methods, including statistical procedures and analyses, see Table 2.4.

Five types of IMs are distinguished: action, reflection, protest, reconceptualization, and performing change IMs. These five types are clearly defined and corroborated with literal excerpts. The methodological procedures for conducting IMCS involve data analysis by two raters who are blind with respect to the outcome (therapy success) of the cases under analysis. Rater training involves the identification of the on-set and off-set of IM as well as categorizing them into the five types. The authors developed a coding training protocol, which when followed successfully results in Cohen's κ of at least .75. So, IMCS fulfills requirements for consistency and transparency.

Linguistic markers which can provide an avenue for the rule-based approach for the five categories are not clearly given in the codebook, but in the theoretical background description, some of the main concepts are discerned with a keen eye to grammar and word use, as can be seen in the distinction between an internal account of oneself ("*My problem is that I have a low motivation to study*") and an intentional state formulation ("*My problem is that I feel that my parents want me to ...*, or "*I'm not so sure that I want that ...*"). The existence of a reliable coding scheme for the analysis of IMs makes this method suitable for the example-based approach. The presented linguistic markers for identity and agency related concepts show promise for translation into the rule-based approach, see Table 2.4.

Conversation Analysis

Under the umbrella term of CA, we found 5 articles that applied to this method to study TCPR, but –like *discourse analysis*– with great variations in its applica-

tion. Although CA for TCPR tends to be used "*more generally and descriptively at the conversation structure of therapy sessions rather than specifically focusing on the change process*" (Elliott, 2010, p. 129), there are notable exceptions such as Voutilainen et al. (2011). However, the method's reliance on in-depth analysis of change sequences within a single case, together with its highly context-bound nature, contribute to low external validity. The authors' own suggestion that the method can be viewed as a complement to other therapy change process methods such as assimilation analysis supports the impression that there is no specific theory or model guiding the method (Voutilainen et al., 2011). So, while the method seems to be a promising addition to TCPR by offering a method for understanding the interaction between counsellor and client, the internal validity of therapy change (the central concept), is low, see Table 2.4.

Regarding reliability, it should be noted that transparency is high in the sense that literal citations are presented with elaborate argumentation anchored in the text. We could not find one specific codebook, procedure or approach that several authors repeatedly used for TCPR purposes (see Voutilainen et al., 2011, for an overview of CA applied for psychotherapeutic research). Commonly, TCPR-researchers referred to Peräkylä (2012), and Voutilainen et al. (2011), see Table 2.3. All in all, the low validity and reliability (at least concerning consistency) of CA, makes its suitability for the example- and rule-based approach highly unlikely in its current form. However, CA is more generally known as a rigorous method, here evidenced by high transparency and adherence to CA rigorous guidelines for anchoring analysis in actual text (Peräkylä, 2012). This characteristic of CA potentially renders it automatable using a rule-based approach to text mining, however, this potential is relatively low in comparisment to the other three methods, see Table 2.4.

Assimilation of Problematic Experiences Scale

The APES method is –compared to the other three methods– based on the most integrated model of therapy change, and presents a generic model to study change that is common to all psychotherapies. Its external validity is substantiated by a wide range of concepts and phenomena from various therapeutic approaches as diverse as cognitive behavioural therapy and psychodynamic approaches. The central idea is that these diverse approaches share a common component: the assimilation of problematic experiences. As the experiences are assimilated, clients move through predictable stages. The generalizability of the assimilation concept is dealt with in a paragraph dedicated to its generality (Stiles et al., 1990, p. 416). That external validity is an explicit goal of this

method, can also be seen its comparative objective of providing "*a reference point for evaluating the effectiveness of alternative techniques*" (Stiles et al., 1990, p. 411).

Reliability of the method is also mentioned from the outset of the article by emphasizing the aim of developing "*a concise, internally consistent, researchable model*". This ambition is then made concrete in definitions of the main concepts (e.g. schema, problematic experience, assimilation and accommodation), and descriptions of the partly overlapping stages of assimilation (unwanted thoughts, awareness, problem clarification, insight, problem solution). The main concepts vary in the extent to which linguistic markers are specified to identify their occurrence in therapy text. For example "*therapists and clients develop words or phrases that bring to mind a constellation of concepts or memories, that is, re-evoke useful schemata*" is much more abstract than example phrases indicating problematic experience (e.g. "*I don't know what this is about*", "*this is not like me*", or "*I can't stand this about me*"; Stiles et al., 1990, p. 412). Application of the method is illustrated using two session summaries, where in-session citations are used to link actual word use to overarching concepts (e.g. schemata). No session transcripts are provided. We found no public access to the mentioned codebook.

All in all, APES scores highest in terms of external and internal validity, but we could not fully assess its reliability due to lack of publicly available information about a code book or inter-rater reliability results. Therefore, we expect that automation through the example-based approach of APES is possible, providing that highly reliable manually coded data become available, see Table 2.4. The potential for a rule-based approach to APES is dependent on the abstraction level of the constructs to be mined. We would suggest that problem experiences are detectable by text mining methods and techniques on the basis of content and grammar, whereas identification of a higher order concept such as schema is much more difficult to automate.

Narrative Process Coding Scheme

The main concepts in NPCS are narrative disclosure, emotional differentiation and reflexive meaning-making. These three concepts allow the identification of three process modes: storytelling, emotion, and meaning-making/reflection. The internal validity and coherence between these concepts is given explicit attention by the authors, who use the dialectical constructivist model to substantiate the presented coding scheme (Angus & Greenberg, 2011). The application of this method is from the outset of this article presented as specific to emotion-focused therapy, although the underlying dialectical constructivist model of therapeutic change appears to have wider relevance to other kinds of therapy. There-

fore, external validity in terms of generalizability of the method to various types of therapy does not appear to be an objective of the authors.

In contrast, reliability of the coding scheme is made an explicit point of attention throughout the article, starting with "*a two-step procedure that enables researchers to reliably subdivide and characterize therapy session transcripts into topic segments*" (Angus & McLeod, 2004, p. 90). Topic segments are defined by key issue and relational focus. These topic segments are then further divided into the three process modes. Identification of these modes requires interpretation of a text excerpt of "*at least ten transcript lines in length*" on the basis of three interpretive questions: storytelling refers to text addressing the question "*what happened to me*", emotion to the question "*what was felt by the client*", and reflection to "*what does it mean to me now*". The authors report that "*good levels of inter-rater agreement were established*" at the level of topic segments as well as process mode. Literal citations of session excerpts are presented, yet no specific word use indicative of the three modes is mentioned. However, the authors mention that they have recently found problem markers (e.g. same old story, empty story), and change markers (e.g. unexpected outcome stories, healing stories). Notably, these markers are at the abstraction level of stories and not at the level of specific utterances.

In sum, we conclude that NPCS as presented in the key article, is in our view both internally valid and reliable, yet its external validity in terms of applicability beyond emotion-focused therapy of depression is unclear. We consider the feasibility of the example-based approach of NPCS high, because of the existence of a highly reliable coding scheme. Despite the lack of linguistic markers reported for the three modes (storytelling, emotion, reflection), we consider the rule-based approach possible, considering that the three modes correspond closely to some of the widely used LIWC categories (cognition, emotion).

Discussion

Because half of the coding schemes we found were used only once or twice, we have come to the conclusion that TCPR is a fragmented field with few attempts for unifying the field. We included the seven manuals of the the four frequently used TCPR methods in our review: IMCS, NPCS, APES, and CA. The first three methods mainly focused on the improvement of the client, CA focused on the interaction between client and counsellor. CA has a low validity and reliability score, rendering both the example- and rule-based approach highly unlikely in its current form. Incidentally, the CA literature on psychotherapy is perhaps

best considered as a specific application of CA, because CA is –primarily– a methodology for the analysis of social interaction, which has –on occasion– been applied to therapeutic data. We therefore do not include CA in the general reflection on the example- and rule-based approach that we give below. We refer the interested reader to Peräkylä (2012) for a review on specific areas of CA and psychotherapy, to Voutilainen et al. (2011) for an example of how CA can be applied in a TCPR-setting, and to Stivers (2015) for a discussion of CA and coding.

The other three methods are in its present state ready for automation to varying degrees, depending on whether an example- or rule-based approach is adopted. Both approaches require an investment for each of the three methods. If sufficient hand-coded data is available, the example-based approach would be very feasible for all three methods; if not, then researchers considering this approach should reflect on whether investing in (manually) annotated datasets outweighs the costs. The rule-based would then perhaps be more feasible.

Example-based approach

We conclude that the existence of a reliable coding scheme for the correct classification of important moments makes the IMCS suitable for automation based on the example-based approach. For the APES, we estimate that example-based automation of APES is possible, providing that highly reliable manually coded data become available. We consider example-based feasibility of NPCS high, because of the existence of a highly reliable coding scheme.

Rule-based approach

The presented linguistic markers for identity and agency related concepts of the IMCS show promise for translation into the rule-based approach. The potential for a rule-based approach to APES is dependent on the abstraction level of the constructs to be mined. Despite the lack of linguistic markers reported for the three modes (storytelling, emotion, reflection), we consider a rule-based approach possible, considering that the three modes correspond closely to some of the widely used LIWC categories (cognition, emotion).

Strengths and limitations

We aimed to cover the whole TCPR field, but only intended to include the frequently used methods in our review. We started by developing a comprehensive search string, but ensured good coverage of all relevant publications of the TCPR

field by also including suggestions of experts. As our systematic review was designed to let the most important and influential publications surface, we felt that we can indeed give an adequate description of the TCPR field.

We also devoted special care to development of our search string as the TCPR field appeared to be a fragmented field. We included queries that covered the psychological literature broadly, the TCPR literature specifically, and text-based methods exclusively. The combination of these queries were used such that we included the whole TCPR field in the initial phase of our review. Because our search string was able to produce the entries that were suggested by experts, we felt that our search strategy covered the literature sufficiently. We adopted the PRISMA standard to ensure that our search was reported transparent and completely.

To avoid placing emphasis on the lesser used methods we focused our study on the commonly used methods, which spanned half of the literature. In doing so, we feel that we included the methods that with a high to good reliability and validity. These quality criteria can only be established through frequent usage, which mandates a focus on the methods used often enough to rightfully claim sufficient reliability and validity. We also valued repeated use of the method particularly because re-use also indicates applicability to multiple data-types and is proof of external validity.

We noticed that the methods that we included mainly focused on the improvement of the client, rather than methods that were tailored to the counsellor. CA focused on the interaction between client and counsellor, but is in its current form not suitable for automated TCPR research. A reasons why we did not find such methods could be because we excluded scientific books from our search. However, as we reviewed the peer-reviewed literature extensively, publications using methods described in books should have surfaced.

It was our choice to exclude methods that were not often used. However, we would like to encourage the researcher interested in these excluded methods –or a specific method– to use our search string. Researchers interested in the (specific) active change processes itself (rather than the TCPR methods), should review the literature with search queries aimed specifically at these processes. From our search, we conclude that these processes do not arise from generic search queries aimed specifically at therapeutic change processes.

We chose to focus only on the text-based methods to study TCPR because we wanted to establish the potential of text mining. Hence, we focused specifically on texts and we did not include non-verbal and behavioural aspects. These aspects contribute to the therapeutic process; however, as they are not textual, they fell outside of the scope of our review. Explorations of these aspects is

therefore open for other TCPR researchers.

Ethical and privacy concerns remain a practical issue for data sharing among TCPR researchers. To stimulate sharing, future research into high quality data anonymization software is now more important than ever. Another option could be for researchers to collaborate in national or international colloquia, that –under strict rules and in line with legislation and informed consent– allow for datasets to be shared internally.

Final remarks

One of the intriguing questions of psychotherapy is how therapeutic processes bring about therapeutic change. As the amount of therapeutic texts is rising in this digital age, new computational methods and research questions became available. These methods enable the investigation of large bodies of therapeutic texts. To facilitate the transition of TCPR from small(er) to Big data, we discussed potential for automation of the most used TCPR methods.

We hope that our systematic review contributes to the unification of the TCPR field. Determining the commonly used methods helps both aspiring and seasoned TCPR researchers to get some perspective on the different research directions in TCPR. Investigation of the potential for automation (of these methods) could also help researchers to determine whether automation is an interesting research direction. We think that automation is important because the (automatic) detection of change processes in therapeutic texts can stimulate the *what works when for whom* discussion, which is not only relevant for psychotherapy researchers, but also for clinical practitioners.

Establishing the potential for automation through the example- or rule-based approach helps TCPR researchers to decide if their datasets are suitable for automation. It also aids in the consideration whether automation is worthwhile. Automated text-analysis methods could also bridge the quantitative and qualitative disciplines, which are sometimes viewed as two different branches of research. Texts –or other forms of verbatim– are usually considered to be the domain of qualitative research, however, large bodies of text mandate the use of quantitative methods.

We therefore strongly feel that automated methods ultimately hold the potential to accelerate psychotherapy research by enabling the investigation of therapy models within and across treatments, groups and settings. Automated methods can thus help gain insights in –for example– innovative moments, or the development of therapeutic change processes over time. Interpretation of these research findings will always require human interpretation, which inevitably

inspires new research questions.


What
Works
When
for
Whom

What
Works
When
for
Whom

3 Text Mining Research of Psychotherapeutic Processes: A State-of-the-Art Review

Abstract

Computerized methods for analysing language are increasingly used in psychological research. Therapeutic Change Process Research (TCPR) is a promising field of application aimed at investigating the linguistic structure underlying psychotherapeutic processes. We provide a State-of-the-Art review of the research objectives, data characteristics and methods of analysis governing automated text mining approaches to TCPR. Combining a systematic keyword search (PsycINFO, Scopus) with snowballing to ensure coverage of psychological and computer science literature we found 32 English peer-reviewed articles. We identified four streams of research: A) *Change Analysts* analyzing emotion-abstraction patterns over time in psychodynamic therapy; B) *Engineers* upscaling and technically advancing motivational interviewing research; C) *Explorers* developing new, complex text mining techniques for theoretical constructs; D) *Digitals* connecting linguistic process markers to therapy outcome in internet counselling. These streams differ in strategies regarding research objective, approaches to data handling and analytical complexity. After discussing the pros and cons of these strategies, we suggest directions for future research that would contribute to the promise of automated TCPR.

Keywords: Text Mining, Natural Language Processing, Literature Review, Future Directions, Psychotherapy Research, Process-Outcome Research

This chapter has been submitted as: Sools, A. M., Smink, W. A. C., Van der Zwaan, J. M., Schuffelen, P. C. J., Tjong Kim Sang, E., de Vries, B. L., Veldkamp, B. P., & Westerhof, G. J. (2019).

Introduction

Language plays an important role in psychological research and practice. People use language to attribute meaning to themselves and their lives as well as to communicate with others. Psychological research focuses on language to infer mental states like emotions and motivations, as well as to determine attributes of relationships, like empathy or conflict (Elliott, 2012; Gelo & Manzo, 2015; Imel et al., 2015; Pennebaker, Boyd, et al., 2015; Salvatore et al., 2015; Schegloff, 2007; Street et al., 2009; Tausczik & Pennebaker, 2010). In everyday practice psychologists use language to motivate others for behavioural change and to promote mental health and well-being (Miller & Rollnick, 2012). Now that psychologists increasingly use computational methods to automatically analyse language use, insight is needed in the current state of knowledge in the field. Therefore, this article provides a state-of-the-art review of automated methods that have been used to analyse textual data in psychology, with a focus on clinical applications in individual psychotherapy, notably *Therapeutic Change Process Research* (TCPR; Greenberg, 2007; Smink, Fox, et al., 2019).[1] TCPR aims to "identify, describe, explain and predict the effects of the processes that bring about therapeutic change" (Greenberg, 1986, p.4). TCPR represents a sub-field in psychology, with broader relevance for text-based analysis of change over time. Based on the state-of-the-art review, we will also set priorities for future investigation and research (Grant & Booth, 2009) that can assist both aspiring and experienced researchers to develop future research in automated text analysis in psychology.

Text mining in psychology

There are two approaches to the automatic extraction of information from text: Text Mining (TM) and Natural Language Processing (NLP). Both combine techniques and methods from many disciplines, including psychology, linguistics, statistics, computer science, artificial intelligence, information retrieval and data mining. TM has a statistical origin and was originally developed for information retrieval, extraction, and summarization in library collections (Miner et al., 2012). Nowadays, it is used to turn all kinds of structured and semi-structured text into numbers, allowing for the application of computer algorithms for further analysis. NLP brings together computer science and linguistics in developing computational methods to learn, understand, and produce language (Hirschberg & Manning, 2015). Its early applications are as diverse as machine

[1] Smink, Fox, et al. (2019) is chapter 6 of this thesis.

translation, speech recognition, and speech synthesis, while more recently NLP is also applied in the creation of dialogue systems and the study of social media (Hirschberg & Manning, 2015). Whereas not all TM is NLP and vice versa, the two computational approaches have a considerable overlap in studying the structure and meaning of texts. In this article we will use TM and NLP interchangeable, to reflect that our review covers both fields. This will open up NLP research to psychology and psychological TM research to the NLP research community.

Over the last few years, TM and NLP are gaining terrain in psychological research. Over the past decades, people have come to use digital media to express themselves, like in social media messages and Internet blogs or in chats and emails in e-health intervention and Internet therapy. Besides these so-called born digital data, the automatic recognition of characters, handwriting, and speech makes it much easier nowadays to digitize data (cf. C. M. Bishop, 2006). The wealth of data that thus becomes available asks for sophisticated methods of analysis in order to better understand psychological processes (cf. Smink, Fox, et al., 2019).

The focus of TM and NLP research in psychology is primarily on social media data. Examples include language use on Facebook in relation to personality traits and life satisfaction (Kern et al., 2016), the broader digital footprint that people leave on Facebook (including likes) in relation to personality traits (Kosinski et al., 2016), negative emotional content of Twitter posts after different traumatic events (Jones et al., 2016), and the prediction of PTSD using posts on a forum for people seeking mental health care (He et al., 2017).

A scoping review of NLP focused on mental health in non-clinical texts, and showed that common data sources include Twitter, Facebook, blogposts on LiveJournal, and a digitized collection of suicide notes (Calvo et al., 2017). These data sources were then used to measure mood and emotion, and to detect depression or other mental health issues. The different NLP approaches in the review used showed face validity, for example that tweets show a more positive mood in the weekends and during seasons with longer days (e.g., Golder & Macy, 2014). Furthermore, others showed that classifying algorithms are able to identify mental disorders based on texts and self-report measures (He et al., 2012), as well as by correlating texts with other data, for example Twitter mood with country-level surveys on life satisfaction (e.g., Schwartz et al., 2013), or a social media depression index with depression prevalence or prescription rates for antidepressant drugs (Choudhury et al., 2013; van den Eijnden et al., 2016). Besides these analytic issues, the authors identify important challenges in this domain such as the interdisciplinary collaboration between psychologists and

computer scientists, the availability of data as well as ethical issues when data are analysed without explicit informed consent of social media users.

These issues arguably play an even larger role in therapeutic contexts, because of the importance of client confidentiality and anonymity. TM studies in clinical settings can roughly be divided in research on screening and diagnosis (Anderson et al., 1999; Kothalkar et al., 2018; Tanaka et al., 2014) and research on the therapy process. We focus our review on the latter application of TM, because of its promise of radically advancing TCPR. TCPR encompasses both qualitative and quantitative approaches. Qualitative approaches are strong in addressing therapy complexity and achieving high ecological validity yet suffer from small sample sizes due to intensive manual labour involved in coding. Quantitative approaches have the advantage of enabling hypothesis testing in large samples, but have limited ecological validity and typically measure only a limited number of variables (Elliott, 2010).

The promise of TM research is that the advantages of qualitative and quantitative TCPR are combined, resulting in a potential break-through in TCPR by enabling large scale testing of therapy change theories and models with high ecological validity. In addition to the potential for theory testing, the application of TM to TCPR also has great practical potential in terms of the improvement of treatments, and training of psychotherapists. In short, a TM approach to TCPR facilitates opening the black box of therapy and allows the investigation of what works when for whom. To investigate how current applications of TM to TCPR fulfill this promise and what is needed to further develop the field, we conducted a state-of-the-art review of research on automated approaches to understanding the linguistic structure underlying psychotherapeutic processes. Our main research question is: *what kind of data are mined with which methods of analysis and for which purposes*? This leads to the following three sub questions: 1) *what are the research objectives reported in TM research on psychotherapy processes* 2) *what are the characteristics of the data and of data handling used in TM research on psychotherapy processes*; and 3) *what are the characteristics of the methods of analysis used in TM research on psychotherapy processes*?

Introduction to Text Mining and NLP

As it can be difficult for newcomers to obtain a thorough understanding of the techniques and terminology that are commonly used, we provide in this section background information to understand the main TM techniques found in our review. For more detailed information the reader is referred to R. Feldman and

Sanger (2007) and Jurafsky and Martin (2017), or Manning and Schütze (1999). We refer the aspiring practitioner of TM to (Bird et al., 2009; Chen & Wojcik, 2016). Readers already familiar with TM can skip this section and go straight to the method section.

Text Pre-Processing, Feature Extraction, and Feature Selection

As a rule of thumb: the text you start with, is not the one you can analyse. In order to be able to automatically analyse text, it must be transformed into a format that is suitable for automatic processing (Snow et al., 2008). Data preparation generally consists of three steps: text pre-processing, feature extraction, and feature selection (Denny & Spirling, 2018). Preparing the texts for TM is called pre-processing. Pre-processing involves deciding on the major entity of study, the unit of analysis, which could be the text documents (which is usually the case), books, web pages, e-mails, tweets, or other text bearing formats. The fundamental unit of each text is a *word*, but before words can be analysed, the unit of analysis needs to be converted into a '*raw*' *text* format: words without any form of typesetting. Then, researchers have to decide on the *segmentation* of (the texts in) their collection of documents, usually referred to as a *corpus*. Depending on the research question, researchers may want to extract client or counsellor text only, divide chat sessions into consecutive sequences of 250 words, speaking turns, or select tweets only from a specific period of time. After deciding on the unit of analysis, a raw text is created for each unit of analysis.

Subsequently, the text in these documents is normalized. Normalization can involve spelling corrections, spelling normalization (e.g., by removing punctuation from abbreviations), and reducing word inflections by applying stemming or lemmatization. Stemming converts words of a sentence to non-changing portions (e.g., '*translation*', '*translating*', '*translated*', and '*translator*' all result in the stem '*translat*'): the most popular stemming algorithm is Porter (1980). Lemmatization is the process of converting the words of a sentence to its dictionary, or root, form (e.g., the lemma of '*meeting*', is '*meeting*' if the word is a noun, and '*meet*' if the word is a verb).

In order to allow for the extraction of text features, raw text documents are represented early on in the preprocessing by a *feature-vector* (Salton, 1971). In the simplest form, each word is represented as a vector (a 'row of several numbers'), and the elements in the vector (i.e. the numbers in the row) express, for example, how often the word occurs. Through these features, it becomes possible to compare documents. The *tf-idf* weighting scheme is commonly used to do so (Salton & McGill, 1986). It assigns a higher weight to words that occur in some

documents and a lower weight to words that occur in all documents (such as articles and prepositions). Weights adjust the elements of the vector relative to their occurrence, or other criteria set by the researchers.

The similarity between two documents can then be measured through these feature-vectors, for example by calculating the *cosine similarity*. This refers to the 'angle' between feature-vectors; a small angle indicates that the documents are similar, while for dissimilar documents the angle is large. For normalized vectors, cosine similarity is the equivalent to Pearson's correlation coefficient. For example, Althoff et al. (2016) use the *cosine distance* to quantify differences between counselling conversations with a positive and negative outcome. The feature-based vector model of a document is also called a 'bag of words' model, because the order of the words in a document is lost. To also capture (some of) the information carried in the order of the words, it is common to use *n-grams*. An n-gram is a sequence of consecutive words. Sequences up to two or three words (i.e., bigrams and trigrams) are in practice used most often to represent documents.

Other aspects of preprocessing features include *part-of-speech* (POS) tags and parse trees. POS tagging refers to the process of assigning categories to words (e.g., noun, verb, adjective, adverb). A parse tree is representing the syntactic structure of a sentence. Both POS tags and parse trees are abstractions that might capture meaning that is relevant for TM. For example, Pérez-Rosas et al. (2017) show that parse trees of counsellor utterances improve classifier performance for the task of predicting whether an utterance is a reflection or not. The number of extracted features in POS and parse trees procedures can get extremely large (i.e. 'explode') relatively fast. Because not all words are equally meaningful (that is to say, not all features are meaningful), the data preparation phase usually involves *feature selection*, the third step. Feature selection can be as simple as removing frequent or infrequent words or removing stop words. More sophisticated methods, such as dimension reduction techniques like principal components analysis (PCA), are used often (Gibson et al., 2015) but beyond the scope of this paper (cf. Hastie et al., 2016).

Text Mining and Natural Language Processing Techniques

After pre-processing, the TM analysis of language use relevant to the research questions can commence. We will describe the four most common TM techniques in Psychology: dictionary methods, rule-based methods, supervised machine learning, and unsupervised machine learning.

Dictionary-Based methods (DB)

The main assumption behind dictionary-based TM methods is that relevant psychological information is conveyed through a person's word use irrespective of their literal meaning and independent of the context in which they occur (Pennebaker et al., 2003). A dictionary contains words that are indicative of the phenomenon the researcher is interested in and can be used to analyse the content of text (i.e., what is being said), and its style (i.e., how it is being said). The prime example of a dictionary-based approach to TM is the well-established program *Linguistic Inquiry and Word Count* (LIWC, pronounced as the English name 'Luke'; Pennebaker, Boyd, et al., 2015). This is a software program that counts the number of words in multiple psychologically relevant categories and reports proportions of words in each category per text. The most recent version of the English dictionary contains almost 6,400 words in 75 categories. Categories include standard language dimensions (e.g., articles, prepositions, pronouns), psychological processes (e.g., positive and negative emotion), relativity-related words (e.g., time, verb tense, motion, space), and traditional content dimensions (e.g., sex, death, home, occupation). Words were included in the dictionary based on ratings by human judges. LIWC has been applied in hundreds of studies (Tausczik & Pennebaker, 2010), and translations have been made to many different languages. LIWC uses unweighted dictionaries, but dictionaries can be weighted too. As an example, the Weighted Referential Activity Dictionary (WRAD) contains 696 words with associated weights indicating the degree to which language reflects connection to non-verbal experience (Bucci & Maskit, 2006). Weighted dictionaries for larger text units (e.g., sentences or speaking turns) can be estimated by applying functions to the norms of the individual words that occur in the text (Malandrakis et al., 2013), for example by calculating the average across words in a text. Based on their review, Calvo et al. (2017) conclude that although all common feature selection and classification procedures have been applied, LIWC is by far the most widely used (Pennebaker, Boyd, et al., 2015).

Rule-Based methods (RB)

Rule-based TM techniques include *regular expressions*, *decision trees*, and *context-free grammars*. A regular expression is a (word)pattern that is used to detect the presence of specific sequences of characters in a text. Regular expressions can also be used to count the number of occurrences of the pattern in the text, which is typically done after documents are represented as *feature-vectors*. For example, Althoff et al. (2016) use regular expressions to detect whether a counsellor

utterance contains a check question, suicide ideation, appreciation, hedges, or surprise. A decision tree is a set of hierarchically organized classification rules. These rules can be hand-crafted, or automatically derived from data. To classify a piece of text, the tree is traversed from root to leaves. At each node, (an attribute of) the input data is subjected to a test. For each outcome of the test, there is a branch in the tree. The branch corresponding to the outcome of the test is followed until a leaf-node is reached, where the input data is assigned the final decision (classification). For example, Gallo et al. (2015) use a decision tree based on hand-crafted rules to classify sentences as open-ended questions. Last, a context-free grammar is a set of production rules that can describe all possible strings in a language 'free of context'.

Supervised Machine Learning (SML)

Psychologists are –perhaps unknowingly– quite familiar with supervised machine learning techniques. The difference between supervised and unsupervised machine learning is described most clearly by contrasting the two. He et al. (2012) collected self-narratives of people with and without PTSD. Based on these classifications, they then tried to find the words that could differentiate the PTSD-patients from the people without PTSD. Now, suppose that this PTSD/non-PTSD classification was not available. Then, a possibility would be to assess whether groups (of people) can be clustered based on only their word use. The former scenario, where outcomes are available, describes *supervised machine learning*; the latter, where no (outcome)*labels* are available describes *unsupervised machine learning*.

The goal of supervised machine learning is to (automatically) find or approximate a function that 'maps' the input variables (i.e., a feature representation of the text) to an output variable (e.g., whether a treatment was successful). The type of the output variable determines whether a *classification* or *regression* should be used. Classification tasks have categorical output variables (e.g., empathic or not empathic), and regression tasks have continuous output variables (e.g., empathy score). In practice, regression is often used as a method to solve classification tasks, e.g. logistic regression for binary classifications.

Generally, supervised machine learning consists of two phases: *training* and *validation*, which requires a split of the dataset in a training and validation set. In the training phase, the 'optimal' approximation between in- and output is calculated based on the training data. This result is then validated by assessing the predictive performance (usually through a confusion matrix) on the (left-out) validation set. It is beyond the scope of this review to describe the many

different available algorithms for training-validation in further detail: we refer the interested reader to C. M. Bishop (2006) or Hastie et al. (2016).

Unsupervised Machine Learning (UML)

In contrast to SML, unsupervised methods do not assume a predefined set of criteria that is used to characterize aspects of the data. Instead, these unsupervised methods are completely data-driven; their goal is to come up with a meaningful characterization of the data. Two unsupervised methods that are used in psychology are topic modelling and sequence labelling. Topic modelling is a technique to uncover the underlying 'topics' of a document collection. It is based on the assumption that words that often co-occur, belong to the same topic. The most widely used topic modelling method is *Latent Dirichlet Allocation* (LDA; Blei et al., 2003). The LDA, perhaps better known as the *topic model*, is a method to discovering the abstract 'topics' in a corpus by clustering documents based on the similarity of words.

Sequence labelling is the task of assigning categories to a sequence of observations, e.g., assigning codes to a sequence of utterances from a motivational interview (Xiao et al., 2012). The advantage of labelling sequences through *n-grams*, instead of single instances is that the context of the current observation can be taken into account when predicting the label assignment. The label of the current observation often depends on the labels of the surrounding observations. For example, if the previous utterance is a question, it is more likely that the current utterance is an answer to that question than a greeting.

Method

Search strategy

The search strategy typically used for State-of-the-Art reviews aims to be comprehensive, yet does not have to be exhaustive (Grant & Booth, 2009). We aimed to be comprehensive by covering both psychological and computational bodies of literature. Therefore, we searched the literature for relevant publications using two complementary strategies: a *systematic keyword search* and *snowballing*. The main difference between the two search strategies is that a keyword search starts with the many and then reduces to the few most relevant publications based on exclusion criteria, whereas snowballing starts with a few key publications and builds from there a wider set of relevant publications based on inclusion criteria. Snowball sampling refers to an iterative process of identifying

relevant publications from other publications by checking references from the related work section and performing searches to find publications that cite this work. We also aimed for a comprehensive time-frame, to allow the full scope of the field to become visible. This means that we considered earlier work in the field as relevant to the current state-of-the-art. Although technically the earlier work may not represent cutting edge research, theoretically it is remains relevant as interest in TM appears to be on the rise and the (research) potential of early work is still far from exhausted.

Databases

Two scientific databases were used for the systematic keyword search: *PsycINFO* and *Scopus*. With many records centred on *psychology*, the *behavioural* and the *social sciences*, PsycINFO is the database that is most likely to cover relevant Psychological literature. We also decided to use the all-round database Scopus, which comes with the additional benefit that it can retrieve records from psychiatry, medical psychology and other relevant medical records as it also includes the database *MEDLINE* (Falagas et al., 2008). In addition, Scopus includes at least some technologically oriented conferences such as *IEEE*.

Search query

Our search query consisted of two parts. To cover the TM literature, we used terms "*text mining*", "*nlp*", "*natural language processing*", "*computational linguistics*", and "*text analysis*" combined by the OR Boolean operator. The second part covered the psychological aspects and consisted of the terms "*therap**", "*psychotherap**", and "*treatment*", again combined by the OR Boolean operator. These two parts were combined using the AND operator.

Snowballing

We started our snowball-search with Althoff et al. (2016) and Atkins et al. (2014), and Howes et al. (2014). Additionally, we screened the 2014, 2015, 2016, 2017, and 2018 proceedings of the *Computational Linguistics and Clinical Psychology* workshop (dedicated to research on language in clinical psychology), and the proceedings of *INTERSPEECH* (the *Annual Conference of the International Speech Communication Association*). The snowball search was completed by using the 'cited by'-functionality of *Google Scholar* to find references of the selected references, and to screen manually the reference list of found papers for potentially relevant additional references.

Identification and selection of methods and studies

We screened the titles, abstracts, and keywords of the articles resulting from both search strategies based on the following four inclusion criteria: (1) peer-reviewed scientific articles and conference proceedings published in the English language (notably articles about *software* for other languages than English were included); excluded were publications such as commissioned report, organizational project paper, book or book review); (2) studies using one or more TM techniques applied directly to psychotherapy or counselling conversations with single clients led by a professional (excluded were indirect measures such as therapist notes as well as screening/diagnosis of for example PTSS, depression and schizophrenia; group-, family, and relation therapy; manuals containing a description of TM tools without applying them to data); (3) sufficiently detailed description of what the researchers did, e.g. when multiple versions of the same study were available the most detailed journal article was selected; (4) TM-only studies (e.g. multi-modal studies which triangulated text with speech or visual modalities were excluded).

In total, our search query retrieved 2174 articles: 926 articles from PsycINFO, and 1248 results from Scopus. After duplicates were removed, the inclusion criteria were applied first on titles, then abstracts and finally full articles when information in title and abstract was insufficient to determine whether a criterion applied. This resulted in 18 publications. In parallel, an iterative snowball search was conducted by applying the same inclusion criteria as we used for the search strategy in the databases until we gathered a set of 29 publications. Fifteen of these publications overlapped with our query-based search. After combining the results of the systematic keyword search and snowball sampling and reducing duplicates, a total of 32 articles was included in our final selection.

Method of analysis

We extracted information from the abstracts and full texts of the selected articles. Because for State-of-the-Art reviews no formal criteria for assessment exist (Grant & Booth, 2009), we created a codebook based on expert consultation with a multidisciplinary team of experts (psychology, computer science and statistics), for analysing the articles in relation to the three sub questions. Below, we describe the different categories that were coded. The specific codes and abbreviations used can be found in Table 3.1 (the four streams mentioned in Table 3.1 will be explained at the beginning of the result section).

Sub-question 1: research aim

Objectives. The research objective was assessed on the basis of the main and (when applicable) secondary and tertiary objectives as explicated by authors. These could be (a) technical (comparing human with machine coding, comparing machine learning approaches or finding best features for an existing construct) or (b) substantive (operationalizing a new construct, relating linguistic process markers to outcome measures, mapping change over time).

Sub-question 2: data characteristics

Client population refers to the kind of patient populations covered by the dataset. In case diverse samples (e.g. multiple client groups diagnosed with different disorders) were used, then this variety was made explicit with multiple codes for each client group. In case multiple disorders were reported within single clients, then the code co-morbidity was used.

Treatment refers to the kind of psychotherapeutic treatment covered by the dataset used in the study. Codes were assigned to the most frequently used treatments (i.e. motivational interviewing and psychodynamic treatment), and infrequent treatments were clustered into a code 'other'.

Sample size. We decided to use *number of sessions* as indication for sample size, because we found this to be the common denominator used in two types of articles: a) articles adopting a population-based rationale for chosen sample size (e.g. number of clients and/or counsellors) and b) those focused primarily on data structure (e.g. number of sessions, utterances, words or tokens). The latter often lacked information about the number of clients and/or counsellors. In addition to number of sessions, we decided to make a distinction between single (S) and multiple (M) case study designs (respectively those focusing on one client over time versus those comparing data from more than one client and/or counsellor – this could usually be inferred even when the exact number of clients/counsellors was not specified) so that the focus of sample size decisions becomes more distinct. It should be noted that single case study designs sometimes included more total sessions (hence effectively used a larger corpus) than multiple case studies where a targeted selection of sessions per case was analysed. Multiple case study designs were further specified into small, medium and large sample sizes based on the number of sessions that were actually used for analysis in the article. This is relevant to mention, because many studies report having made selections from a larger dataset.

Annotation level details the level of the dataset that is annotated (i.e., provided with labels). Annotation could be based on pre-existing labels or labels created

specifically for the research at hand. The levels are treatment level (e.g. metadata such as client characteristics like gender or outcome measures of overall client improvement), session level (e.g. level of empathy shown by counsellor), and within-session level (e.g. utterance, excerpt or word level annotations). Few studies did not annotate the data, but only used automated analysis.

Sub-question 3: data-analysis

Construct complexity describes the complexity of the constructs that are investigated. A single complex construct such as empathy is operationalized by multiple text features. A single reduced construct breaks a construct into aspects or dimensions (e.g. type of reflection) and only focuses on part of the full construct. Also indicated is whether the linguistic structure of one or multiple psychological constructs were investigated. A last code is assigned when the temporal unfolding of linguistic features over time is the explicit topic of investigation (rather than aggregating over treatments or sessions).

The *unit of analysis* can be client and/or counsellor and/or the relation between client and counsellor. In cases when annotated datasets are re-used with labels for more than one unit of analysis, the actual unit of analysis chosen for the study at hand is coded.

Reproducibility of the pre-processing procedure is indicative of the reliability (transparency and trackability) of the procedure. The codes here cover the range from articles where the reported procedure is unspecified to limited, partially, and completely reproducible (e.g. available on GitHub or similar).

Text mining or NLP technique refers to which of the (combination of) four main TM techniques described in the background section in this article were used in the article: DB (dictionary based), RB (rule based), SML (supervised machine learning), UML (unsupervised machine learning).

Under *Software* we differentiated between the most frequently used software (LIWC, Therapeutic Cycles Model; TCM) other existing software, and self-developed software. Some articles did not explicitly mention the kind of software used.

The *statistical and technical methods of analysis* are clustered into codes indicating the extent to which these methods (a) belong to the standard repertoire of psychologists, (b) need additional basic programming expertise and software or (c) need high level programming expertise or specialized software. The latter two categories are currently not commonly used by psychologists, although they may be standard in computational linguistics. This classification gives insight in the kind of analytical repertoire psychotherapy researchers need to conduct TM.

Table 3.1: Codebook with abbreviated codes, code descriptions and frequency counts per category for all four streams.

Abbreviation	*Description*	N_A	N_B	N_C	N_D	N
Research aim (technical, n = 30; or substansive, n = 31)						
CHM	Comparing Human-Machine	1	8	3	0	12
CML	Comparing Machine Learning approaches	1	3	0	0	4
FBF	Finding Best (text)Features	1	10	1	2	14
RPO	Relate linguistic Process markers to (treatment/session level) outcome	4	2	4	2	12
OC	Operationalize (propose new) Construct	1	3	5	1	10
MCT	Mapping Change over Time	6	0	3	0	9
Client population						
SA	Substance Abuse	0	13	0	0	13
DA	Depression and/or Anxiety	6	0	1	2	9
PD	Personality Disorder	4	0	2	0	6
ED	Eating Disorder	0	0	0	0	0
O	Other	4	2	2	0	8
U	Unspecified	1	1	5	0	7
Co	(Co-morbidity) Multiple disorders within single patients are reported	2	0	0	0	2
Treatment						
PA	(brief) Psychodynamic or psychoanalytic treatment	6	1	5	0	12
MI	Motivational Interviewing	0	13	0	0	13
CBT	Cognitive Behavioral Therapy	1	2	0	1	4
U	Unspecified	1	0	1	0	2
O	Other, e.g. client-centred therapy, crisis, counseling, group therapy, prevention	0	1	2	1	4

Table 3.1 continued from previous page

Abbreviation	*Description*	N_A	N_B	N_C	N_D	N
Sample size						
S	One client followed over time	3	0	4	0	7
MS	*Small* 2 or more clients, total session number up to 100	2	2	1	0	5
MM	*Medium* total session up 100 and 800	3	10	2	0	15
ML	*Large* total session number over 800	0	2	0	2	4
Annotation level						
T	Treatment level annotated or meta-data at corpus-level	5	2	3	1	11
S	Session Level annotated	4	10	1	1	16
WS	Within-Session Level annotated	1	10	6	1	18
None	Only automated analysis was performed	1	0	1	0	2
Construct complexity						
Single complex	Single construct defined by multiple features	4	6	6	0	16
Single reduced	Single construct partial or reduced aspect or dimension	1	2	0	0	3
Multiple	Multiple constructs	2	5	1	2	10
Temp	Attention to temporal unfolding	6	10	0	0	16
Unspecified	Not specified	0	1	0	0	1
Unit of analysis						
Cl	Client language	8	6	8	2	24
Co	Counselor language	5	14	4	1	24
R	Relation between client and counselor language	4	2	1	0	7

Table 3.1 continued from previous page

Abbreviation	*Description*	N_A	N_B	N_C	N_D	N
Reproducibility of pre-processing						
Completely	Software is specified and pre-processing steps are described so that the procedure is repeatable and transparent (e.g. available on GitHUB or similar)	0	0	0	0	0
P	Software is specified and an overview is given of the pre-processing steps	3	2	1	2	8
L	No software is specified but an overview is given of the pre-processing steps	4	12	6	0	22
U	No software is specified and no overview of pre-processing steps is provided, but reference to description in previous article may be made	1	0	1	0	2
Text mining / NLP techniques						
DB	Dictionary Based	8	4	5	2	19
RB	Rule Based	1	8	5	1	15
SML	Supervised Machine-Learning	0	6	2	2	10
UML	Unsupervised Machine-Learning	0	2	1	1	4
Software						
TCM	Therapeutic Cycles Model	7	0	0	0	7
LIWC	Linguistic Inquiry Word Count	1	5	1	2	9
OSS	Other existing Software Specified (name in brackets)	1	2	5	0	8
DOS	Developed Own Software	0	0	3	1	4
NM	Not mentioned	0	8	0	0	8

Table 3.1 continued from previous page

Abbreviation	*Description*	N_A	N_B	N_C	N_D	N
Statistical and technical methods						
Basic (GC, PL, MP)	Methods and techniques that are standard available in psychology (Group Comparison, Predict Label, Model Performance)	8	19	9	0	36
Intermediate (LDA, NG, ROC)	Methods and techniques that require additional programming skills and/or software (Latent Dirichlet Allocation, N-Grams, ROC-curves)	0	7	1	0	8
Advanced (NN, SVM, HMM)	Methods and techniques that require high level technical programming expertise and/or specialized software (Neural Networks, Support Vector Machines, Hidden Markov Models)	0	8	1	2	11

Results

We found a total of 32 English peer-reviewed articles with the first publications emerging in the 1990s [1, 2, 23; numbers refer to references in Table 3.2]. Based on analysis of the research aims, data characteristics and data analysis methods reported in these articles, we identified four streams of research (see for details Table 3.1 and 3.2). After characterizing each stream separately in the following sections, we will give an overview of the field as a whole. We assigned multiple sub codes to each article, therefore, the frequencies mentioned in Table 3.1 can have a higher total than the 32 articles that entered the review.

A. Change analysts: Analyzing emotion-abstraction patterns over time in psychodynamic therapy

Stream A ($n = 8$) represents one of the historically first groups (and collaborations of this group with members of researchers in other countries), based in Germany, that pursued automated text analysis of psychotherapy text with the first articles appearing in 1996. Drawing on research on cycles of therapeutic change,

the TCM (*Therapeutic Change Model*) software was created. This model aims at analysing emotion-abstraction patterns over time. The four patterns that have been identified (relaxing, experiencing, connecting and reflecting) are considered to represent change cycles that apply across treatments and psychopathologies, and are hypothesized to meaningfully connect to treatment outcomes.

The research aims in this stream are predominantly substantive in nature, with as common thread a focus on mapping change over time [1-6; number refers to the numbers in Table 3.2] and connecting process to outcome [1, 4, 5, 7]. Only two articles report explicit technical aims [2 and 8]. The studies in this stream mainly focus on depression, anxiety and personality disorder patients undergoing psychodynamic or psycho-analytic treatment with one exception [6] applying TCM to Cognitive Behavioral Therapy (and one article [3] with unspecified treatment). The datasets under investigation vary and may be derived from widely researched databases such as the Penn Psychotherapy Study [1] or databases derived from practice (e.g. forensic psychiatry [3]). Typically, a purposive selection from larger datasets is made in two steps. First, case selection takes place of either exemplary cases to detect for example significant change moments [4, 6, 8] or contrasting cases comparing most and least improved clients [1, 2, 5, 7]. Second, an additional selection of sessions out of a larger number of sessions per client is made, apparently to limit time-consuming manual annotation. Three single case study designs [3, 4 and 6] and five multiple case study designs [1, 2, 5, 7, 8] have a medium sample size of on average 153 annotated sessions ($SD = 202$). Most frequently used are annotations at treatment level ($n = 4$) and annotations at session level ($n = 4$). Within session level annotation ($n = 1$) and no annotation ($n =1$) are rare in this stream. Treatment level annotations use a combination of clinical ratings, self-report and questionnaires for depressive symptoms and global functioning. Annotations at session and within session level cover measurements for session quality, therapeutic alliance and helpful aspects of therapy.

The main construct in this stream, consisting of the cycles in abstraction-emotion language, is complex with its attention to temporal complexity (e.g. how language develops over time both within sessions and over sessions). The unit of analysis is predominantly client language ($n = 8$), but also counsellor language ($n = 5$) and the relation between counsellor and client ($n = 4$) are often included. The researchers in this stream pay attention to making the preprocessing procedure reproducible (limited and partial in 7 out of 8 articles), without however making the procedure available on a software developer platform. Reliance on TCM Software with accompanying dictionaries is distinctive for this stream, with one study making adaptations to the dictionary [7] and

one adding LIWC [8]. With DB as main text mining technique and group comparison as preferred method of analysis, this stream is technically basic. What is compared differs between studies, e.g. comparison between most and least improved clients [1, 5, 7], between psychotherapies [2], between client groups [3], or between different parts of the therapy [4, 6, 8]. For this comparison, researchers often adopt a mixed method approach in which machine analysis and human interpretation go hand in hand.

Overall, this stream combines technically basic analysis with theoretically complex analysis to study a potentially widely relevant construct (therapeutic cycles) with a good basis in the TCPR tradition. The studies in this stream show that therapeutic cycles can indeed be meaningfully identified in a variety of client populations, but the evidence for trans-treatment validity is currently limited (mostly psychodynamic). The step-wise reduction of initial larger datasets results in actual medium sample sizes that could theoretically have been larger. Datasets are typically not re-used, which limits possibilities for gradually building larger annotated datasets. More advanced machine learning approaches are not employed in this stream, which may be due to the initial head start of the pioneering group who started the TCM research.

B. Engineers: Upscaling and technically advancing motivational interviewing research

The second stream $(n = 14)$ is geographically predominantly located in the US, where researchers from various universities joined forces to use automated text analysis for upscaling motivational interviewing research. At the basis of their collaboration (with the first article in 2010), lies a corpus consisting of up to six datasets from previous research on counsellor skills in Motivational Interviewing (MI) targeting substance abusers $(n = 13)$. The MI corpus is sometimes complemented with additional datasets such as the Switchboard DAMSL (*Dialog Act Markup in Several Layers*) dataset [12] and the Alexanderstreet corpus [14], presumably to increase sample size. Two studies are boundary cases of this stream: One involves MI research by authors apparently unconnected to the main group of collaborating authors and using a different corpus [21], the other with overlap in collaborating team members but concerning CBT rather than MI [22]. The MI corpus is partially annotated (and appears to have received additional annotations along the way) predominantly at within session level $(n = 10)$ and session level $(n = 8)$ with MISC (*Motivational Interviewing Skill Code*; Miller & Mount, 2001) and MITI (*Motivational Interviewing Treatment Integrity*; Moyers et al., 2005) coding schemes for assessing counsellor fidelity to motivational in-

terviewing. With an average of 372 sessions used for analysis, the typical sample size in this stream is medium with two outliers [12, 14] showing large sample sizes of over 1300 and 1500 sessions. Overall, random selections of annotated sections of larger samples are made for actual analysis, hence effectively reducing potentially larger sample sizes available.

In this stream, technical research aims ($n = 21$) clearly outweigh substantial research aims ($n = 5$). Over time, there appears to be a gradual shift from studies aimed at showing that automated analysis is a good alternative to human coding ($n = 8$) to studies aimed at optimizing the operationalization of the main constructs (i.e. counsellor fidelity and empathy) through finding best text features ($n = 10$) and comparing computer models ($n = 3$). The optimization process appears to strike a balance between accuracy and efficiency (e.g. optimizing prediction based on reduced construct operationalizations) versus validity and nuance (e.g. seeking linguistically more fine-grained operationalizations). Examples of efficiency-driven studies are those investigating the predictive power of one aspect or dimension of counsellor skills such as reflections [9, 17]. Examples of nuance are studies investigating the sequence of utterances within sessions [10]. Most studies show construct complexity in either looking for single complex constructs ($n = 6$), temporality ($n = 10$) or multiple constructs ($n = 5$). When attention is paid to temporality, it is usually about sequences within sessions, rather than development over sessions. All studies in this stream focus on counsellor language ($n = 14$), although there is some attention to client language ($n = 6$) and the relation between counsellor and client ($n = 2$).

All 14 studies in this stream provide limited information relevant to reproducing pre-processing steps, which may be due to the fact that many publications are conference proceedings with limited space. We did not find evidence that the procedure is made available on a software developer platform like GitHub. Similarly, there is also a lack of specification about the programming software that is used ($n = 8$) and of the 7 cases in which the software is mentioned this mainly concerns LIWC ($n = 5$) and not the software used for the other TM techniques. Researchers in this stream typically use multiple TM methods and techniques within single studies with RB ($n = 8$) and SML ($n = 6$) outweighing DB ($n = 4$) and UML ($n = 2$). The methods of analysis also show a wide variety with predominantly basic methods ($n = 19$) followed by advanced methods ($n = 8$) and intermediate methods ($n = 7$). Although group comparison is also a popular method of analysis in this stream ($n = 7$), model performance is the most frequently used method ($n = 8$). Neural networks are the most frequently employed advanced method ($n = 4$), suggesting a move toward unsupervised learning to circumvent labor-intensive annotation

procedures.

Overall, we get the impression of a technically oriented group with good access to computer engineering expertise, who are dedicated to upscaling MI research and experimenting with new technological solutions. While we observe that researchers in this stream take an effort to increase sample sizes by combining datasets (with upscaling analysis as explicit goal), we also noticed that selections of annotated sessions prevent full-scale use of available datasets. Experiments in this stream point towards unsupervised machine learning as upcoming solution for this annotation problem.

C. Explorers: Developing new, theoretically complex text mining approaches

This stream ($n = 8$) is a mixed bag of researchers from various geographical locations who have in common that they each work toward developing TM approaches for constructs that are not covered with existing approaches. These are either theoretically or linguistically complex constructs. Four of the articles [23-26] in this stream concern operationalizations of affect (e.g. valence, arousal; emotional access), which can be seen as alternative or complements to LIWC. More advanced approaches to language concern speaking or language style [23, 24], Discourse Flow Analysis [27] and context-sensitive meanings [28]. The last two articles respectively offer a set of linguistic features to analyse body boundary [29] and repair language [30]. Most articles are published in the 2010s with earlier exceptions in 1999 [23] and 2007 [24] that however do not seem to have received a follow-up, at least not in TCPR applications.

Substantive research aims ($n = 12$) dominate this stream with the proposition of new constructs as its main feature ($n = 5$). There is some attention to technical aims, notably comparison of human and machine interpretation ($n = 3$). The datasets used vary per study and could be either derived from practice or involve counsellor training data or data collected for effect studies. Psychodynamic treatment is the preferred choice for the affect studies, while the remaining studies concern other therapies, e.g. client centred therapy [29]. Strikingly, the client population is in more than half of the studies unspecified ($n = 5$). This stream is characterized by small sample sizes (average of 76 sessions, $SD = 80$) with relatively few cases (e.g. clients) as shown also in the frequent choice for single case studies ($n = 4$). Within-session level annotations are most frequently employed ($n = 6$) as compared to treatment level ($n = 3$) and session level ($n = 1$), hence showing an interest in micro-level processes.

The constructs under investigation show high complexity regarding multiple

aspects of a phenomenon such as language style [24] attended to (e.g. referential activity, affect and reflection, disfluency, non-verbalized vocalized responses). All studies focus on client language, with some attending also to either counsellor language ($n = 4$) or the relation between client and counsellor ($n = 1$). The pre-processing procedure is relatively limited, at least when compared to more elaborate description of mostly newly developed software dedicated to the new construct that is proposed, e.g. CALAS [23], DAAP with WRAD dictionary [24, 26], Turkish language affect analysis system [25], DFA [27], ACASM [28], body type dictionary [29]. This new software mostly adopts DB ($n = 5$) and RB ($n = 5$) text mining techniques, with some attention to SML ($n = 2$) and UML ($n = 1$). Finally, most studies ($n = 6$) in this stream use one or more basic methods of analysis (notably group comparison and model performance). Two studies use intermediate and/or advanced methods, e.g. neural networks [27], Support Vector Machines and Latent Dirichlet Allocation [30].

All in all, the constructs in this stream show theoretical and linguistic complexity. For this purpose, new dictionaries are developed and complemented with RB approaches. Technically, this stream appears to employ relatively basic methods and techniques, with some use of more advanced ones. Sample size is relatively small, and re-use of software and datasets seems limited. This reflects that these are newcomers in the field for which the future will tell whose work may eventually develop into their own streams and research groups.

D. Digitals: Connecting process to outcome in internet counselling

Despite the small number of studies ($n = 2$), we felt that a separate cluster should be dedicated to an upcoming stream of researchers investigating chat- and SMS-texting-based counselling. Because only two studies are in this stream, the following characterization is rather tentative. Characteristic of the born digital data studies in this stream is that they have large sample sizes available (up to 15,555 sessions). Alongside the substantive research aim focused on connecting linguistic process markers to outcome, both studies also report the technical aim of finding best features. To connect process to outcome, the first study [31] has labels at treatment level (meta-data for gender, age, etc.) and session level pre- and post-test labels for depressive and anxious symptoms. The second study [32] has session level binary labels (improved and not improved). Howes et al. (2014) analyse discussion topics and sentiment to show that written therapy can be compared with face-to-face data. However, they conclude that patient progress can only be "*captured by finer-grained lexical features suggesting that as-*

pects of style or dialogue structure may also be important." Similarly, the findings of Althoff et al. (2016) indicate that a more comprehensive model achieved the best performance. Both studies derive construct complexity from including multiple text features that go beyond mere word counts to include for example ambiguity and creativity [32]. Although both studies use LIWC as starting point for their analysis, they note its limitations and use additional TM techniques, e.g. SML [31 and 32], RB and UML [32]. Moreover, they both use an advanced method of analysis (i.e. Hidden Markov Model), making these studies to be among the technically most advanced of the field.

Table 3.2: Characteristics of the 32 articles included in our review, divided in four streams.

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
Stream A. Emotion-abstraction patterns in single cases of psychodynamic therapy ($N_A = 8$) 153 sessions ($SD = 202$)											
1. Mergenthaler (1996)	MCT, RPO, OC	DA, PD, O	PA	MM; 108	T + S	Single complex + temp: Therapeutic Cycles (emotion abstraction patterns: relaxing, experiencing, connecting, reflecting)	Cl + Co	L	DB	TCM	basic (GC)
2. Mergenthaler & Kachele (1996)	CML, MCT	U, DA	PA	MS; 64	None	Multiple + temp: formal, grammatical, and substantial text features	Cl + Co + R	P	DB, RB	TCM	basic (GC)
3. Päfffflin et al. (2005)	MCT	Co	U	S; 62	S	Multiple + temp: Emotional experience, cognitive mastery, connecting	Cl	P	DB	TCM	basic (GC)
4. McCarthy et al. (2011)	MCT, RPO	Co	PA	S, 16	S	Single complex + temp: Therapeutic cycles (see under 1)	Cl + Co + R	U	DB	TCM	basic (GC)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
5. McCarthy et al. (2014)	RPO, MCT	PD, DA	PA	MM; 400	T	Single complex + temp: Therapeutic cycles (see under 1)	Cl + Co + R	L	DB	TCM	basic (GC)
6. Sassaroli et al. (2014)	MCT	DA	CBT	S; 10	WS	Temp: CBT interventions (assessing, disputing, and reframing biased beliefs)	Cl + Co + R	P	DB	TCM	basic (GC)
7. Boldrini et al. (2017)	RPO	DA, PD, O	PA	MM; 540	T	Single reduced: Mentalization measured by reflective functioning	Cl	L	DB	OSS (CRF)	basic (GC)
8. McCarthy et al. (2017)	FBF, CHM	PD, DA	PA	MS; 20	T + S	Single complex: Significant events (emotion, reflection language and alliance strengthening)	Cl	L	DB	LIWC, TCM	basic (GC)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
Stream B. Upscaling motivational interviewing research $(N_B = 14)$ 372 sessions $(SD = 463)$											
9. Can et al. (2012)	CHM	SA	MI	MS; 57	WS	Single reduced + temp: Counsellor reflection	Co	L	SML	NM	basic (MP), intermediate (NG), advanced (HMM)
10. Xiao et al. (2012)	FBF, CHM	SA	MI	MM; 116	T + S + WS	Single complex + temp: Counsellor empathy	Co	L	SML	NM	basic (PL), intermediate (PL), advanced (SVM)
11. Atkins et al. (2014)	CHM, FBF	SA	MI	MM; 148	S + WS	Single complex: Counsellor fidelity	Co	L	DB, UML	NM	basic (GC, MP), intermediate (LDA, ROC)
12. Can et al. (2015)	CHM, FBF	SA	MI	ML; 1303	S + WS	Multiple + temp: Empathy and spirit	Co	L	SML	OSS (ASR lattices)	basic (MP, PL), intermediate (NG)
13. Gibson et al. (2015)	FBF, CHM	SA	MI	MM; 148	S + WS	Single complex: Counsellor empathy	Co	L	DB, RB	LIWC, OSS (Semval-2014)	basic (GC, MP)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
14. Imel et al. (2015)	OC	SA, O	CBT, PA, MI, CCT, other	ML; 1553	T + S	Multiple: Sensible topics, therapist statements, therapeutic relationship	Cl + Co	L	UML	NM	basic (GC), intermediate (LDA)
15. Lord et al. (2015)	FBF	SA	MI	MM; 112	S	Single complex + temp: Counsellor empathy (synchrony in language style, therapist reflections)	Co + R	L	DB, RB	LIWC	basic (GC, PL)
16. Tanana et al. (2015)	CML, CHM	SA	MI	MM; 356	WS	Multiple + temp: Counsellor fidelity, client change talk, client sustain talk	Cl + Co	L	RB, SML	NM	basic (PL), advanced (NN)
17. Can et al. (2016)	CHM, FBF	SA	MI	MS; 57	WS	Single reduced + temp: Counsellor fidelity (reflections only)	Co	L	RB, SML	NM	intermediate (ROC), advanced (HMM)
18. Gibson et al. (2016)	FBF, RPO	SA	MI	MM; 348	S + WS	Single complex + temp: Counsellor empathy	Co	L	RB	NM	advanced (NN)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
19. Tanana et al. (2016)	CML, CHM	SA	MI	MM; 341	S + WS	Multiple + temp: Counsellor fidelity; client change talk, client sustain talk	Cl + Co	L	RB	LIWC	basic (MP), advanced (NN)
20. Gibson et al. (2017)	FBF, CML	SA	MI	MM; 337	WS	Unspecified	Cl + Co	P	RB	NM	basic (GC, MP)
21. Pérez-Rosas et al. (2017)	OC, FBF, RPO	SA, O	MI	MM; 277	S + WS	Multiple + temp: Counsellor empathy; participant engagement; participant mimicry patterns; discussion topics	Cl + Co + R	P	RB	LIWC	basic (GC, MP)
22. Flemotomos et al. (2018)	OC, FBF	U	CBT	MM; 92	S + WS	Single complex + temp: CBT Competency	Cl + Co	L	DB, SML	LIWC	basic (GC, MP, PL), advanced (NN)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
Stream C. Developing theoretically and/or linguistically complex new TM techniques, tools and software ($N_C = 8$) 76 sessions ($SD = 80$)											
Affect (*CALAS* / *LAVAS* and *DAAP*)											
23. Anderson et al. (1999)	RPO	PD, U	PA	MS; 32	WS	Multiple: Affect, speaking style, stylistic complexity	Cl + Co	L	RB	DOC (CALAS)	basic (GC)
24. Bucci & Bernard (2007)	CHM	U	PA	S; 16	S + WS	Single complex: language style (referential activity, affect and reflection, disfluency, nonverbalized vocalized responses)	Cl + Co	L	DB	OSS (DAAP with WRAD)	basic (GC)
25. Halfon et al. (2016)	OC, CHM, MCT	DA, O	PA (play therapy)	MM; 120	T + WS	Single complex: Affect (valence and arousal)	Cl	L	RB	DOS (Turkish affect analysis)	basic (GC, MP)
26. Halfon et al. (2016)	OC, MCT	U	PA	S; 30	WS	Single complex: Emotional access (verb repetition + Referential Activity)	Cl	L	DB, RB	OSS (DAAP with WRAD)	basic (GC, MP)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
Discourse Flow Analysis (*DFA*)											
27. Nitti et al. (2010)	OC, MCT, RPO	U	PA	S; 43	None	Single complex: Patient therapist communicative flow (amount of activity, superorder nodes, connectivity)	Cl + Co + R	L	DB, RB, SML	OSS (DFA)	advanced (NN)
Automated Co-occurrence Analysis for Semantic Mapping (*ACASM*)											
28. Salvatore et al. (2012)	OC, CHM	PD	O	S; 48	WS	Context-sensitive meanings	Cl	U	RB, UML	OSS (ACASM)	basic (GC)
Body Boundary											
29. Cariola (2015)	OC, RPO	U	O (client centered therapy)	MM; 240	T	Single complex: Body imagery linguistic variables (i.e. anger, body, plural and singular pronouns)	Cl	L	DB	LIWC, DOS (Body type)	basic (GC, MP)

Table 3.2 continued from previous page

Reference	*Aim*	*Client*	*Treatment*	*Sample size*	*Annotation level*	*Construct(s) complexity*	*Unit of analysis*	*Pre-proc.*	*TM*	*Software*	*Statistical methods*
Repair											
30. Howes et al. (2012)	RPO, FBF	O	U	U	T + WS	Single complex: Repair	Cl + Co	P	DB, SML	OSS (Weka toolkit, SVM Light)	intermediate (LDA), advanced (SVM)
Stream D. Connecting process to outcome in internet counseling ($N_D = 2$)											
31. Howes et al. (2014)	RPO, FBF	DA	CBT	ML; 882	T + WS	Multiple: discussion topics and sentiment, patient progress	Cl	P	DB, SML	LIWC, DOS (Stanford classifier)	advanced (HMM)
32. Althoff et al. (2016)	RPO, FBF, OC	DA	O (crisis intervention)	ML; 15555	S	Multiple: Counsellor adaptability; Ambiguity; Creativity	Cl + Co	P	DB, RB, SML, UML	LIWC	advanced (HMM)

Overview of the field across the four streams

Overall, substantial aims ($n = 31$) and technical aims ($n = 30$) both occur equally, with stream B taking the bulk of the technical aims (21 out of 30), stream A and C focusing on substantive aims, and stream D aiming to achieve substantive aims by technical advancements. Datasets used in stream A and B have rather homogenous population and treatment characteristics (e.g. depression and anxiety clients undergoing psychodynamic treatment in stream A; substance abusers in motivational interviewing data in stream B). Interestingly, CBT, by far the most used and most widely researched psychotherapeutic treatment in the Western world, is clearly under-represented ($n = 4$) in the studies included in this review.

Stream D has the largest sample size (<1000), followed by medium sample sizes of over a hundred sessions in stream B (372) and A (153), and small sample size (<100) in stream C. In all but stream D, actual sample sizes are a result of a selection process in which parts of a larger dataset are sampled based on availability of annotations (or space within the project to create additional annotations). About half of the studies use a multiple case study design of medium sample size, with stream A, C and D typically reporting sample size (also) at population level (number of clients and/or counsellors), and stream B generally privileging data structure to indicate sample size (number of words, tokens, sessions). Within session annotations ($n = 20$) outweigh session ($n = 16$) and treatment level ($n = 10$) annotations, but this distribution is skewed (stream B outnumbers all other streams). Only two studies report no use of annotations, making the use of annotated data for TM the preferred option in this field.

Except for some studies aiming for construct simplicity (explaining more with less features), most studies prefer to add complexity either by adding text features ($n = 16$), by including multiple constructs ($n = 10$), or by seeking theoretically or linguistically more complex constructs (with stream C as prime example). Half of the studies ($n = 16$), all in streams A and B, pay attention to temporal complexity. This is in stream B typically mapping change in language use within sessions, and in stream A both within and across selected sessions. We did not find any studies mapping change over the course of a complete therapy trajectory of one or more clients. Overall, there is much less attention to the relation between client and counsellor language ($n = 7$) than to language use of client ($n = 24$) and counsellor ($n = 24$) separately (with counsellor language with $n = 14$ distinctive for stream B). Most studies do not provide information about the pre-processing steps with a level of detail that allows for immediate reproduction ($n = 24$), and we found no studies reporting to have made avail-

able pre-processing steps on a software platform such as GitHub. This may be explained by the change to open science standards that become more prevalent nowadays when compared to the time when the majority of these studies were published. DB is the most frequently used TM technique ($n = 19$) often as part of the TCM model in stream A. Of the other dictionaries, LIWC is the most frequently used one ($n = 9$). Second in line are RB approaches ($n = 15$), followed by SML ($n = 10$) and UML ($n = 4$). Group comparison ($n = 21$) is the preferred (basic) method of analysis across all streams. Stream B and stream D show the greatest variety in TM methods used within single studies with model performance ($n = 8$) as the most popular one. These streams also adopt the technically most advanced methods of analysis.

Discussion

In this literature review we investigated the state-of-the-art in the field of automated text-based approaches to psychotherapy research. For this purpose, we performed a State-of-the- Art review to investigate which kind of data are mined with which TM methods and techniques and for which purpose. We observed four main research streams in the field (e.g., A. *Change Analysts*, B. *Engineers*, C. *Explorers*, and D. *Digitals*). These four streams differ in decisions and considerations regarding three closely related methodological aspects, which correspond to the three sub research questions in this article: 1) the research aim, the way they handle 2) data issues and 3) analytical issues. In Figure 3.1, an overview of these decisions is presented that can help both aspiring and experienced researchers to inform their study designs.

In response to the first research question about the research aim, we conclude that the division between researchers prioritizing substantive versus technical research aims indicates different developmental phases in the field. The difference between streams A and C suggest subsequent phases of software development and wider implementation. After initial development of a new model (e.g. TCM), stream A has invested in building a knowledge base supporting the external validity of the model across client populations (but not yet across treatments). Stream C researchers are in the initial development phase of software for linguistically and theoretically complex constructs, and the future will tell whether their models and techniques will be taken up by other researchers to become established techniques. The relative priority given to technically advancing motivational interviewing research in stream B may reflect the need to first develop better TM approaches and acquire larger annotated datasets before

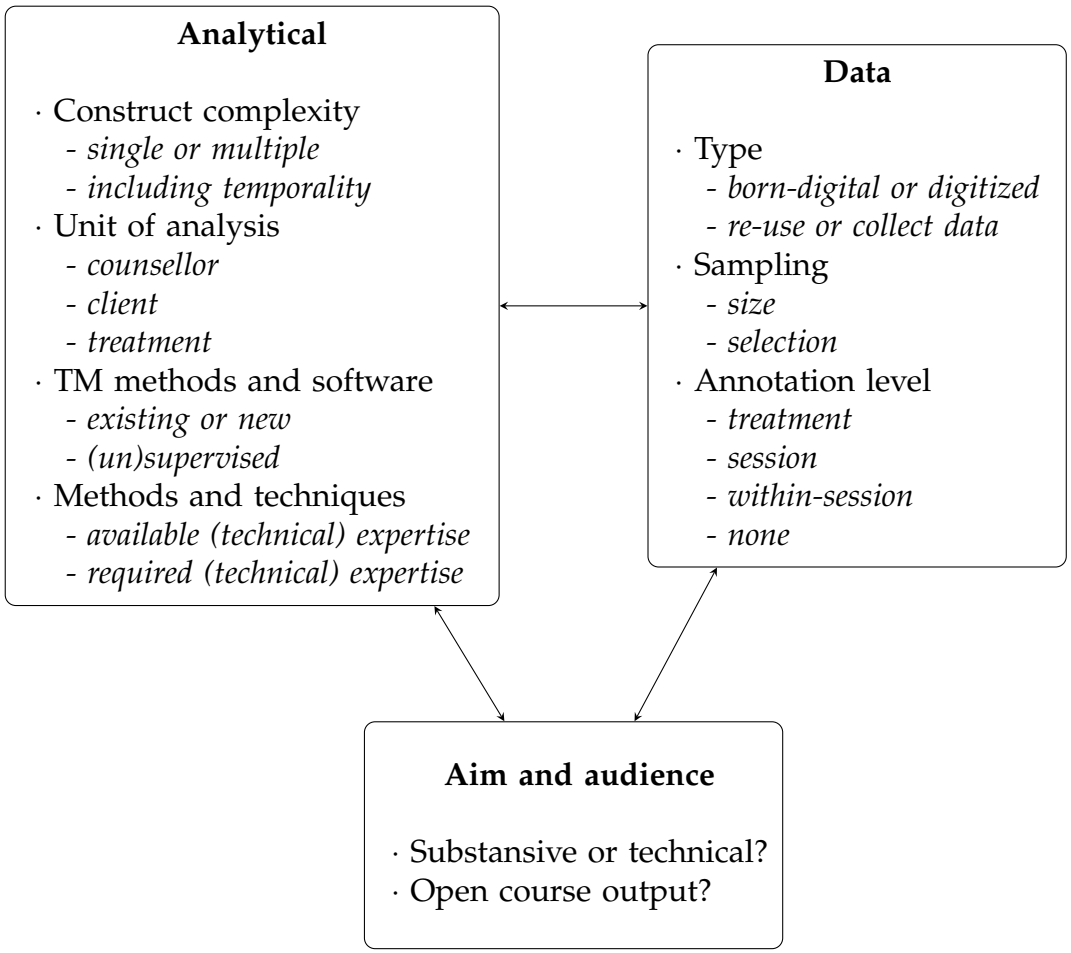


Figure 3.1: Overview of decisions involved in designing Text Mining research.

actual testing of hypotheses comes into view. In stream D the requirement of large annotated datasets is fulfilled, hence laying the foundation for large scale research on the connection between process and outcome. However, the relative weight attributed to substantive and technical research aims appears to be also influenced by the disciplinary background and composition of the research teams (e.g. computer sciences or psychology). Therefore, the issue of choice at formulating the research objective appears to be partly pragmatically informed (e.g. based on available technical expertise and dataset specifications) and partly on (discipline-specific) substantial interest.

In response to research question two, data characteristics, we found two main factors affecting sample size decisions: 1) availability of annotated data, and 2) role of human interpretation. Even when large (video- or audiotaped) datasets are available, often only a selection is used to circumvent labor intensive transcription and annotation tasks. While selective use of data is understandable and often necessary within the scope of a research project, this practice also hinders using the full potential of machine learning, which benefits from large datasets.

We found some inventive strategies to solve the annotation problem in stream B that might be useful for other researchers as well. They re-used datasets and gradually built a larger annotated corpus, they combined datasets, looked for more efficient coding schemes, and adopted unsupervised machine learning to circumvent labor intensive manual annotation. However, these solutions may not always be possible or desirable. Re-use of datasets for example may be difficult because of privacy and informed consent issues. Moreover, the practice of selecting sessions may be informed by the important role attributed to human interpretation by many TCPR researchers. This could explain why fully automated and unsupervised machine learning is not more frequently adopted.

However, instead of limiting the number of sessions in advance based on theory or logic (e.g. the first or third session), we recommend making informed selections based on human interpretation of language use detected automatically over the whole corpus. This procedure may allow to find new, unexpected findings of what counts as significant moments in therapy, and to detect predecessors of these moments. In addition, the rise in internet-based treatments has the exciting prospect of using born-digital data on a wider scale. These datasets are among the largest currently available in the studies included in this review. A big advantage of born-digital data is that time-intensive transcription of face-to-face conversations is not needed. In this respect, automatic transcription software also offers possibilities, but we did not encounter this application in our review. With only 2 studies (distinctive for stream D) using born-digital data this potential is underutilized. Notably, the small number of born-digital studies was also due to the exclusion criteria (two articles of group-wise internet-based counselling and one about a medical application were excluded). We would argue that beyond pragmatic choices for born digital conversations, TM research opens a host of psychologically relevant (new) research questions such as what the therapeutic relation looks like in web-based counselling.

In answer to the third research question about data analysis methods, we conclude that we see a trend towards seeking either theoretical, linguistic, or technical complexity in each stream. We also noticed lack of crossovers between streams, except for stream D, indicating fragmentation of the field. Stream A researchers started out with a theoretically complex construct, using basic technical and statistical methods to show the external validity of this construct. Generalizability of the therapeutic cycles to different client populations is found, yet generalizability to psychotherapy treatments other than psychodynamic therapy needs to be tested in future research. This stream is comprehensive in including all units of analysis (client, counsellor, relationship and temporality). Even more comprehensive future research could entail including complete client tra-

jectories (not selections), and comparing more clients to build a more robust evidence base. Moreover, including software developed for linguistically and theoretically more complex constructs in stream C could enrich the TCM model. This is important, because the initiative to develop TCPR-specific software (in stream A as well as C) shows that LIWC is not the primary choice in the field of TCPR although it is sometimes used as starting point or benchmark ($n = 9$ out of 32 articles). We also recommend looking at neighboring disciplines (e.g. computational linguistics) for potentially relevant NLP approaches. Examples of computational linguistics relevant for psychology and TCPR include for example empathy in call center conversations (Podsiadlowski et al., 2013), embodied meaning in child language (J. Feldman & Narayanan, 2004), embodied emotions in historical texts (van der Zwaan et al., 2015), computational modeling of narrative (Mani, 2012).

Stream B focuses on technically advancing the field to find more fine-grained ways of detecting (mainly) counsellor fidelity, with mixed results. Completer models, in particular when features are enriched with contextual information, tend to yield better performance than simpler ones. However, interpretation can be difficult in for example deep learning approaches that are becoming more popular (although we only encountered neural networks). Therefore, investing in optimizing seems worthwhile, providing that is recognized that the optimization process takes time, resources and patience. The choice for empathy and reflection as main constructs may stretch the potential relevance of the approach in this stream beyond MI to other treatments, but this may require better embedding in psychological theory to be successful. In case the trend to include client talk and synchrony between client and counsellor language [15, 21] continues, this will increase the attractiveness of stream B for researchers outside the MI field. Finally, connecting (within) session level processes to treatment outcomes, similar to the studies in stream D, could also be an important next step to test when and how counsellor fidelity matters. In light of current open science requirements (for example motivated by Woelfle et al., 2011), we recommend to make software and code used to pre-process and analyse the data available on an online repository, such as GitHub.

Conclusion

Overall, we conclude that TM research in the field of psychotherapeutic change is a relatively small yet promising field of pioneers. A potential explanation is that the complexity of therapeutic change processes requires going beyond

basic TM methods and techniques. Evidence for this idea can be found in the search for increased theoretical, linguistic and technical complexity. The four streams we identified seem to specialize in either theoretically, linguistically or technically advancing the field. The streams also differ in a focus on comprehensiveness (client, counsellor, relationship, with eye for temporality) or upscaling TCPR research. A big obstacle lies in the lack of structured, complete, and comparable datasets (Veldkamp, 2018): most data are unstructured text with no clearly identifiable structure and/or predefined data model (He, 2013, p.1). Current strategies to reduce data handling complexity reflect the preparatory status of the field. With expected technological developments (automatic speech detection) and the increase in born-digital data, the conditions for upscaling TCPR research are expected to improve over the next few years. However, to speed up the preparatory process investment in the creation of sharable datasets with comprehensive annotation systems will be needed. When sharing of datasets is difficult (for example due to privacy issues which are abound in this field), then at least re-use of TM methods and techniques between researchers from the four streams would be highly recommended. Finally, interdisciplinary knowledge exchange would be vital to bring the ambition of a unified field capable of answering comprehensive research questions closer into view.

What
Works
When
for
Whom

4 The Automation and Explication of Research Methods: Understanding their Interplay through a Framework, with Therapeutic Change Process Research as an Use-Case

Abstract

There are differences between the research goals, models and methods of the developers of text mining applications and the users of these text mining applications. To get insight in these differences, we introduce the automation-explication framework and show why disciplinary preferences make it difficult to automate qualitative research methods with text mining. As an use-case, we classify the four streams of text mining Therapeutic Change Process Research (TCPR). The framework shows that the rule-based approach prevails for methods aimed at explanation, whereas the example-based approach predominates for methods aimed at improving the accuracy of predictions. By visually showing the disciplinary differences, we hope that the framework can bring users and developers of text mining methods closer together (or that it at least becomes easier to communicate over the borders of disciplines).

Keywords: automation-explication framework, Therapeutic Change Process Research (TCPR), text mining, automation, explication

This chapter has been submitted as: Smink, W. A. C., Sools, A. M., Tjong Kim Sang, E., Veldkamp, B. P., & Westerhof, G. J. (2020).

Introduction

Even though there are many parallels between text mining and qualitative research methods, not all qualitative methods are –in their current form– well-equipped to analyse the large volumes of data that are available nowadays (Ho Yu et al., 2011). Even though the 'talking cure' lies at the heart of psychotherapy, none of the methods for *Therapeutic Change Process Research* (TCPR) are –in their current form– suitable for the analysis of 'Big data-sets' (Smink, Sools, van der Zwaan, et al., 2019).[1] In the current paper we argue that the cause lies (in part) in the difference between the goals, methods and models of the *developers* of (new) text mining applications (who are usually mainly computer scientists), and the *users* of these text mining applications (such as psychologists). We identify two trade-offs that can give insight in these differences: the orientation of the *explication*, and the method of *automation*. Understanding how users and developers differ with respect to these trade-offs could stimulate the interdisciplinary collaboration that psychotherapy research requires in this day and age.

We aim to facilitate the multidisciplinary nature of psychotherapy research in two ways. First, we introduce the *automation-explication* framework (which links the two trade-offs), and second we demonstrate how the framework can be applied. To do so, we discuss how automation and explication differ between the streams of text mining TCPR that were identified by Sools et al. (2019).[1] By emphasizing what approaches can be used for automation (we make a distinction between *rule-* and *example*-based approaches) it becomes clear there is a difference in how conclusions can be explicated (we distinguish between models aimed at *accuracy* and *explication*).

Towards that end, we organised this article as follows. In the remainder of this introduction, we will discuss how the two sides of both trade-offs co-exist. In the *framework*-section, we discuss the four quadrants of the framework. We then discuss the use-case of TCPR in the *application*-section. We address the general and TCPR-specific implications of the framework in the *implications*-section.

Explication

Because the co-existence of *accuracy* and *explanation* is relevant for all research disciplines that work with data, (some form of) the trade-off is addressed by machine learning, data science and statistics (Donoho, 2017). We mainly relied

[1] Smink, Sools, van der Zwaan, et al. (2019) and Sools et al. (2019) are chapter 2 and 3 of this thesis.

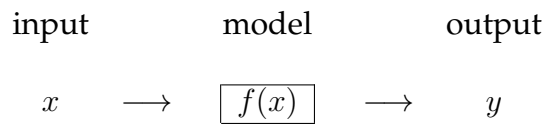


Figure 4.1: The (statistical) model $f(x)$ maps the independent variable x to the dependent variable y.

on Breiman (2001b) and Shmueli (2010) to cover the statistical science; for machine learning and data science we used *arXiv.org* publications (Duval, 2019). We will first discuss the trade-off broadly, and then discuss the relevance in the field of TCPR. We like to stress that accuracy and explanation are not mutually exclusive, but researchers are implicitly confronted with a choice between these concepts when models need to be optimized. For example, for basic and very simple cases, it is possible to obtain 100% accuracy, but 100% explication is not an objective that makes much sense (i.e. what is the difference between 95% and 100% explanation). Nevertheless, in practice, it is common to distinguish between accuracy and explanation as the two orientations of research goals.

The explication trade-off can be understood through the three elements in Figure 4.1: the independent variable x (also known as the *input* variable), the dependent variable y (the *output*), and the model $f(x)$ that associates x and y. The trade-off arises because the predictive accuracy of the model $f(x)$ improves at the expense of its explanational value (Kuhn & Johnson, 2018). To phrase it differently, the difference between *accuracy* and *explanation* lies in the complexity that is allowed for $f(x)$: modelling non-linear (or non-smooth) relations between x and y results in more accurate predictions for y. However, as a result, the complexity of $f(x)$ increases, which reduces the explanational value of $f(x)$.

In the literature, it is not uncommon that different names are used for the concepts of accuracy and explanation. Improving the accuracy is also referred to as *predictive*, *algorithmic*, or *optimized* modelling. Explanation is also known as *interpretative*, *informative*, *transparent*, or *simple* modelling. We prefer the terms *accuracy* and *explanation* because they inherently express the two different goals of data analysis, and organised these concepts as the trade-off in Figure 4.2. We propose to use the term *explication* to address the trade-off itself, as the term expresses that there are two approaches to explicate a statistical model (see Figure 4.1 and 4.2).

explanation ‹- -› accuracy

Figure 4.2: The explication-axis describes whether the model aims for *explanation* or *accuracy*.

The explication trade-off

So, the complexity that is allowed for $f(x)$ (see Figure 4.1) differs for both sides of the trade-off in Figure 4.2. It turns out that, in practice, different research disciplines favour different sides of the trade-off. Applied researchers (such as social, economical or medical scientists) mainly value explanation: they perform data analysis to increase their (theoretical) knowledge about how the input variable x and output variable y are associated in $f(x)$. That becomes difficult if $f(x)$ is too complex and non-explainable to a (human) researcher. To give an example, a model that simply lists all the patients at risk of developing a certain mental health issue is not helpful without offering insight in the specific risk factors.

There are also disciplines that do not emphasize the explanation of $f(x)$: knowing how x and y are related in $f(x)$ is of secondary importance, as long as $f(x)$ returns good predictions for y. Because model complexity is not-restricted, the model is 'allowed' to use everything for good predictions, so it can 'choose' to employ complex, non-smooth, and non-linear functions for $f(x)$. As a result, the predictions for y become more accurate, at the price that $f(x)$ becomes a black-box (i.e. is not explainable). A translation algorithm is a good example: correct translations are more important than an understanding of *how* this result is achieved. Certainly, an understanding of what makes translations 'good' is helpful, but the goal of the application is prediction (rather than explanation).

So, the practical difference between the two approaches (i.e. two sides of Figure 4.2) is what exactly is required of the model $f(x)$. Some disciplines mainly value accuracy, others prefer explanation. Because explanation is the result of restricting the model complexity, accuracy and explanation are often in conflict as accuracy requires increased complexity (which is why presenting these concepts as a trade-off in Figure 4.2 is appropriate). Arguably, this is also one of the reasons why an interdisciplinary collaboration is not straightforward to organise: the *type of preferred conclusions* differ between research disciplines. In the *application*-section, we will show that technically orientated disciplines appear to pursue accuracy-related models more often than domain-specific research like psychologists (and that this relation is the other way around for explanation).

	Predicted positive	*Predicted* negative	
Observed positive	True Positive (TP)	False Negative (FN) Type II error	**Sensitivity** $\frac{TP}{TP+FN}$
Observed negative	False Positive (FP) Type I error	True Negative (TN)	**Specificity** $\frac{TN}{TP+FP}$
	Precision $\frac{TP}{TP+FP}$	**Neg. Pred. value** $\frac{TN}{TN+FN}$	

Figure 4.3: The confusion matrix.
Note. Often used measure for the overall model performance are the *accuracy* (calculated as the sum of all the correct classifications, the true positives and true negatives, divided by the sum of all measures in the Figure, and the F_1-*score*, calculated as the (harmonic) mean between the precision and sensitivity. *Neg. Pred.* is an abbreviation for *negative predictive*.

Type of conclusions

With the risk of over-generalizing, those valuing accuracy are –arguably– less sensitive to the explanation-issue because they rely on the confusion matrix (see Figure 4.3) to assess model performance. Complex (and therefore less explainable) models are more accurate and lead to a 'better' confusion matrix (i.e. a higher *accuracy* and F_1-score, see Figure 4.3). However, the matrix itself does not address model complexity, which is why model evaluation based on the confusion matrix alone could make the *developers* of new text mining methods less sensitive to the needs and preferences of the *users* of their models. These users typically prefer explainable (and therefore less complex) models, which also means that the confusion matrix is going to consist out of suboptimal performance metrics (which is not the goal of model optimization).

Even when model complexity is considered, it turns out to be difficult to determine what an acceptable level of model complexity is. What could be an insignificant increase of complexity for a developer, could mean the difference between a model that does or does not have value for an user. On the other hand, the small increase of 4% accuracy (from 95% to 99%) could mean the difference between an application that *does* or *does not* have value for practical use (for example, voice-controlled applications only have value if they perform nearly perfect). All in all, being unaware of these subtle but important differences leads to explanation-preferences that do not translate well across the borders of

disciplines.

Relevance for TCPR

As we intend to go beyond a theoretical discussion alone, we also discuss explanation in the context of TCPR, a field dedicated to find an answer to the question *what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of *circumstances* (Norcross & Wampold, 2011; Paul, 1967; Tasca et al., 2015). Arguably, this *What Works When from Whom* (WWWW) question lies at the heart of psychotherapy research, which makes TCPR a relevant field for many researchers (hence our choice for this specific use-case).

Two recent literature reviews of TCPR present an overview of the field. Smink, Sools, van der Zwaan, et al. (2019) concluded that there was potential for automation for the majority of the often-used qualitative TCPR methods (which is the topic of the next section), but "*due to their dependence on human interpretation, these methods are limited in analysing the large bodies of text that are nowadays available, limiting their use to small scale research.*" In other words, Smink, Sools, van der Zwaan, et al. (2019) note the importance of explication, but limit their conclusions to automation (the topic of the next section) alone.

The other review, conducted by Sools et al. (2019), distinguished four streams of automated TCPR research: *change analysts*, *engineers*, *explorers*, and *digitals*. Sools et al. (2019) concluded that the streams differ with respect to the research objectives, methods and data characteristics. However, Sools et al. (2019) do not address the difference in explication between streams explicitly, and only cover automation briefly. We will discuss the study of Sools et al. (2019) in detail in the application- and implication-section. Based on these two reviews, we conclude that the concept of explication is not yet explored for TCPR.

We also chose this field because we expected that it would have long-standing discussion with respect to explication. We expected to find different approaches –with different historical roots– to the WWWW question. In line with the behaviouristic tradition, we expected to find some who argued that an accurate prediction tells 'everything' about behaviour (i.e. 'accurately predicting behaviour = understanding behaviour'). We also expected that others took the position that the mechanisms that underlie therapy are more important for understanding change, and thus favoured explainable models (i.e. explainable models = understanding behaviour). The two reviews that we use in the current article do not discuss these questions in detail, and –to our surprise– this discussion was absent in the literature.

Automation

Due to their dependence on human interpretation, qualitative research methods are limited in analysing the large bodies of text that are nowadays available, constraining their use to smaller samples. To adequately understand why automation of qualitative methods is not straightforward through a text mining approach (aside from difference in explication), we (also) need to consider the *automation*-trade-off. *Automation* is when a piece of machinery, technology, or an algorithm performs a process or procedure without –or with minimal– human assistance. We use the verb *automating* to describe the process of replacing human labour by an automatic process (i.e. 'outsourcing the labour to technology').

Text mining is a framework of automated methods for the automatic processing of natural language. There is more than one approach for automatic text analysis, in our context of text mining –and similar to Smink, Sools, van der Zwaan, et al. (2019)– we will discuss and highlight the importance of *rule*- and *example*-based automation. Because rule-based automation is best understood in its historical context, we shortly discuss the history of the field first (Jurafsky & Martin, 2014).

Rule-based models

The roots of text mining are closely intertwined with those of *Natural Language Processing*: NLP originated in computer science, and one of its earliest applications was deciphering the encrypted Morse-code messages sent by Nazi-Germany. After the Second World War, others also tried to model natural language through *rule-based* language models (Johnson, 2009; Jurafsky & Martin, 2014). The work of these pioneers emphasized the core of rule-based models: input variable x is mapped to output y through some function $f(x)$ (Figure 4.1 is also helpful here). Rule-based language models describe a set of models that explicitly define the relation between x and y through a set of unambiguous rules for $f(x)$. Because of the explicit definition of rules, these models can typically be well-understood by humans: even though the relation between x and y is 'complex' (e.g. $y = e^x$), rules are concise, explainable, and therefore easy to understand (Weiss & Indurkhya, 1995).

Typical rule-based language models include (Brzozowski, 1964; Chomsky, 1956; Cremers & Ginsburg, 1975; McNaughton & Yamada, 1960; Porter, 1980): *regular expressions* (a sequence of characters that define a search pattern), *context free grammars* (a set of rules used to generate all possible word combinations based on an input string), and *stemming* (reducing inflected (or sometimes derived) words to their word stem; i.e. an algorithm tasked with the goal of finding

the verbs ending in '*-ing*'). There are some models, such as *named entity recognition* (a model that locates *named entity mentions* such as person names, organizations, locations, medical codes, time expressions, and quantities), that nowadays have rule- and example-based alternatives, but that are in practice still mainly rule-based (Marrero et al., 2013). For rule-based approaches, each word in the text –x in Figure 4.1– is represented by a set of features (Salton & McGill, 1986), these features are then compared against rules for $f(x)$, and a specific rule is applied if a match between x and one of the features is found (so that x is mapped onto y).

For the majority of practical purposes, rule-based models did not age that well, with some exceptions in educational settings (Muñoz & Montoyo, 2007; Yoon et al., 2008). Practical applications of language models require the specification of an unworkably large set of rules, which turned out to be infeasible given realistic time- and budgetary constraints. Although rule-based models are concise and explainable, they are typically plagued by a low *sensitivity* (Sofaer et al., 2019): these models only work for (very) specific cases (like *regular expressions*, *context free grammars*, and *stemming*) that all need to be pre-defined explicitly and well-tested before they can be used in practice.

Sensitivity refers to a trait of the confusion matrix in Figure 4.3. The *precision* (also from the confusion matrix) of rule-based models is relatively high for these specific use-cases, but not prone to erroneous input, which is almost always a part of human produced language. Familiarity with the confusion matrix can help researchers understand why rule-based models are less popular nowadays, as these models cannot be scaled easily beyond the specific application for which they were developed, which greatly limits their value to their specific use-case (Himmel et al., 2009). This is also the main reason why the history of text mining and NLP are marked by a gradual shift away from rule-based models.

Over the years, scientists from different fields (such as computer science and electrical engineering) began experimenting with language models that were not based on rules, but that 'learned' to model language based on 'raw' examples from texts (Jurafsky & Martin, 2014). Around 1990, this led to what many refer to now as a *statistical revolution* (Martinez & Martinez, 2015): example-based *machine learning* models became more prominently featured in text mining than rule-based models (Johnson, 2009; Manning & Schütze, 1999).

Example-based models

Around the 90s, computational resources and the availability of data both greatly increased (for example, the *Linguistics Data Consortium* became accessible for ev-

eryone), making way for example-based models, which require a lot of data and computational power (Hilbert & López, 2011; Jurafsky & Martin, 2014; Liberman, 2002). It turned out that the probabilistic data-driven models from statistics and machine learning are better equipped for yielding a high precision and –when sufficient training data is available– a similar recall as rule-based models (Sofaer et al., 2019). In about the span of a decade, these *example-based* models completely took over the field (Martinez & Martinez, 2015).

To sharpen the contrast with rule-based models we propose to call these models *example-based*, instead of 'statistical' or 'machine learning' models (Smink, Sools, van der Zwaan, et al., 2019). This name also expresses the core of these models (M. Bates, 1995): they rely on statistical inference to automatically learn the 'rules' of language by analysing large text corpora containing typical real-world examples of language (instead of analysing language through explicitly and pre-defined rules). Example-based models 'learn' the function $f(x)$ by providing a (machine) learning algorithm with specific examples of how input x and output y should be associated. Based on these examples, the model then searches for an form of $f(x)$ that mimics the these associations.

The result is that example-based models typically focus on learning the common cases of a text, as $f(x)$ is most 'familiar' with these cases. When writing rules for rule-based models it is not always obvious where the (first) effort should be directed at (Franke et al., 2016), because there is an extremely large number of candidate rules available for each language (making the question of where to start difficult). Example-based models have several other advantages: they handle exceptions to (grammar-)rules (or spelling) well (without specifying all these exceptions), they can model words and verbatim that were not featured in the training set, and they are relatively robust against erroneous input. Another advantage is that the accuracy of example-based models is simply a function of the amount of input data: training $f(x)$ with more examples of how x and y are associated leads to better model performance (Banko & Brill, 2001; Martinez, 2010). There are more advantages, but we focused on those with the largest impact on practical use (Jurafsky & Martin, 2017; Manning & Schütze, 1999). The disadvantage of example-based models is that the larger they get, the more they start to behave as accurate models in Figure 4.2 (and as a result it requires considerably more effort to make these models explainable). In practice this means that if the number of examples becomes (quite) large, the explication of the models is affected (hence the reason why we propose to organise these concepts in a trade-off, see the application-section). It turns out to be difficult to create models that both perform well in terms of accuracy, and that at the same time explainable.

The prominence of example-based automation also had another effect on the field: it also meant that Chomsky's theories of linguistics became of less importance (Norvig, 2017). The theoretical underpinnings of Chomsky discourage the use of '*corpus linguistics*'. "*One of the big insights of the scientific revolution, of modern science, at least since the seventeenth century... is that arrangement of data isn't going to get you anywhere. You have to ask probing questions of nature. That's what is called experimentation, and then you may get some answers that mean something. Otherwise you just get junk*" (Aarts, 2001). Indeed, Chomsky was a fierce critic of the implicit assumption that example-based models make: he believed that simply modelling more data about language did not lead to a better understanding of language (Sanyal, 2009).

Different methods of automation

Because TCPR has a long standing tradition in the analysis of studying the linguistic 'products' of therapy (e.g., homework exercises, diaries, transcripts, p. 303, 392), there are many different qualitative TCPR methods available (Gelo et al., 2015). As there is more (text) data available nowadays, it becomes even more important to consider automated approaches to these traditional methods (Sools et al., 2019). In the previous section, we made a case for a distinction between rule- and example-based approaches to automation by discussing their differences and historical origins. Although perhaps less relevant for practical purposes in text mining, a distinction between rule- and example-based approaches is important when assessing the potential of automation of qualitative methods in general, and for TCPR specifically. These approaches express how a researcher can automate a research method in practice: they can either *specify a large set of rules*, or they can present *annotated examples of the in- and output variables*.

In the application-section, we will show that there is not much difference between the fields that develop new text mining applications and fields that apply text mining applications, but there is a difference between what is considered to be a '*simple*' method. For example, the *Linguistic Inquiry and Word Count* software of Pennebaker, Boyd, et al. (2015) could be considered as a standard example-based application of text mining with low accuracy, but is of great use and importance for applied fields such as psychology also because of its great applicational value. To reflect that research methods can rely on both approaches (such as named entity recognition), the rule- and example-based approach form the two ends of the automation-axis in Figure 4.4. The automation axis indicates that function $f(x)$ from Figure 4.1 can be approximated by either '*giving examples*' or by '*defining rules*'.

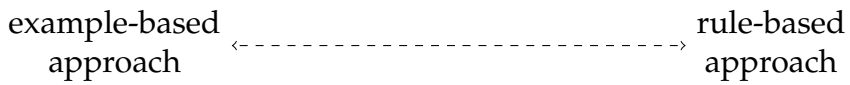


Figure 4.4: Similar to Smink, Sools, van der Zwaan, et al. (2019, doi:10.1371/journal.pone.0225703.g001), automation can be seen as an unidimensional trait which can be achieved either through an example-based approach, or a rule-based approach, with hybrid methods in-between.

Goals

We distinguished the differences in explication (accuracy or explanation), and automation (rule- and example-based approaches). In the next section, we use these two concepts as axes in a framework. In doing so, we explicitly assume that these two concepts are independent of each other, and that both accuracy and explanation can be achieved through a rule- or example-based approach. We also introduce the framework itself in the next section; in the application-section we demonstrate how the framework can be used by applying it to TCPR.

The framework

With explication and automation addressed in Figure 4.2 and 4.4 respectively, we can now introduce the *automation-explication* framework. Under the framework presented in Figure 4.5, the explication-trade-off (accuracy and explanation) coexists with two different approaches to automation (example- and ruled-based). It follows directly from the framework that different approaches to automation can be associated with different types of explication. This coexistence lies at the heart of the framework: when considering how to automate qualitative research methods, researchers are confronted with a choice between *accurate* or *explainable* models, guided by *example*- or *rule*-based automation.

The four quadrants of the automation-explication framework

These trade-offs differ between the four quadrants of the framework (i to iv in Figure 4.5). We discuss the framework through these four quadrants and give some examples from TCPR methods. We have to note that –as with any theoretical model– it is possible to have different opinions on how certain methods and models could (and should) be classified. We used our own experiences working as an interdisciplinary team as the main method of analysis for the framework. As a result, the strengths and limitations of what we present here is what is 'common sense' for us, guided by the question '*what does this mean in practice for an interdisciplinary approach*'.

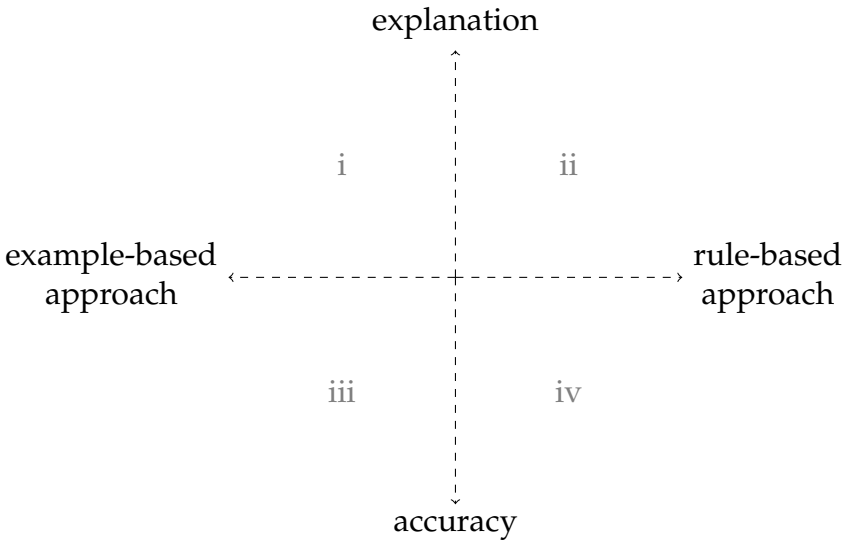


Figure 4.5: The automation-explication framework is a balance of automation approaches (*rule-* and *example*-based) and the kind of explication (*explainable* or *accurate* models). These two axes distinguish four quadrants (i – iv).

Quadrant i

The core of methods in quadrant i is that they rely on the *example-based approach* with the goal of advancing *explanation*. Their key trait is that they are *dictionary-based* (Pfäfflin et al., 2005), and often used to summarize certain psychological characteristics that are expressed in text (e.g., *emotional experience*, *cognitive mastery*, *connecting*). As the name implies, dictionary-based methods typically comprise of a 'dictionary', a list –or several lists– of words that reflect some sort of domain-specific ontology of a psychological category (Hoogendoorn et al., 2017). For example, *anxiety* could be operationalized by words such *scared*, *panic*, and *afraid*. The underlying assumption is that relevant psychological information is conveyed through word use (Sools et al., 2019).

To assess this information, a dictionary contains a (finite) list of words (i.e. *examples*) that should be recognized by the method, which are used as an indication of the *theoretical* phenomenon that the researcher is interested in. The LIWC program by Pennebaker, Boyd, et al. (2015) is the prime example of a dictionary-based method. LIWC counts words into (pre-defined) meaningful categories and allows for predictions of behavioural outcomes, as well as identifying words that reflect the underlying psychological state of a person (Sools et al., 2019). For example, McCarthy et al. (2017) used LIWC to identify significant events in treatment (such as *emotion*, *reflection*, *language* and *alliance strengthening*).

Strengths

The main strength of quadrant i methods is that they are able to explore theoretically relevant text-characteristics. Dictionary-based methods are often used for psychotherapy research (Sools et al., 2019). Because dictionary-based methods can assess a large number of texts simultaneously (Smink, Sools, van der Zwaan, et al., 2019), they opened up new possibilities of text research. Back et al. (2011) used LIWC analyse the emotional words used in the 422,502 text pager messages sent on September 11, 2001.

Potential pitfalls

One drawback of quadrant i methods is that they are dependent on the written and standard version of a language, and therefore insensitive for the differences in spoken language or dialect (Kamps et al., 2004). LIWC is available in ten different languages: English, Chinese, Arabic, Spanish, Dutch, French, German, Italian, Russian, and Turkish (Pennebaker, Boyd, et al., 2015). Although it is always possible to 'simply' translate a dictionary from one language to another, it is not guaranteed that this results in a useful translation. Unfortunately, the language-dependency limits the usability of the method outside that language. Often, quadrant i methods are available in English and sometimes also in a second language, frequently the mother-tongue of the research group, such as Italian (Sassaroli et al., 2014).

Dictionary-based methods require only the specific words for the psychological construct of interest, but it can be difficult to determine which words are relevant without a theoretical overview or advice from experts. If a researcher wants to explore a construct beyond the available examples, it i difficult to do so without text-data on the topic. For example-based approaches in general it is not enough to have only a few examples, since constructs need to be supported by sufficient data to constitute a dictionary.

Quadrant ii

The core of methods in quadrant ii is that they rely on the *rule-based approach* to advance *explanation*. These methods are used to extract text-characteristics through rules based on theoretical knowledge of the construct. In doing so, these methods provide a basis for further development or adjustment of theory. Regular expressions and decision trees are typical rule-based methods. Althoff et al. (2016) used regular expressions to detect whether a counsellor utterance contained a *check question*, *suicide ideation*, *appreciation*, *hedge*, or *surprise*. An-

derson et al. (1999) discriminated between 'good' and 'poor' counsellors, and from high affect segments they extracted and labelled verbs as *basic*, *experiencer*, or *benefactive* based on several rules they developed. Gallo et al. (2015) used a decision tree based on hand-crafted rules to classify sentences as open-ended questions.

Strengths

The main advantage of these is ease of implementation. It is straightforward to specify a rule-based model, and put it to work quickly to see what kind of results it returns. Rule-based approaches are therefore also a good first exploratory step. If the research objective is achieved through only a few rules, the rule-based approach of quadrant ii often suffices.

Potential pitfalls

If a researcher intends to use a rule-based model outside of the specific example that it was developed for, then usually a large set of (additional) rules need to be specified. This means that the rule-based methods of quadrant ii typically involve a lot of manual work which require a deep (and usually a graduate level of) understanding of the specific field. Also, complex theoretical constructs are not easily extracted through only a few rules. It is therefore not surprising that some traditional methods from this quadrant are less used nowadays (but quadrant ii is where automated text analysis originated).

Quadrant iii

The core of methods in quadrant iii is that they rely on the *example-based approach* with the goal of improving *accuracy* (with little to no attempt to make these methods explainable). The origin of these methods lie in *problem solving*, such as speech, image and handwriting recognition, non-linear time series prediction, and prediction in financial markets (Breiman, 2001b). For example, Nitti et al. (2010) designed a neural network to identify patterns of functioning of the discursive network and to verify the clinical validity of these patterns in terms of their association with specific phases of the psychotherapeutic process.

Strengths

The main advantage of these methods is that they have good predictive properties, and can work relatively autonomously. We limit the discussion between

supervised and *unsupervised* methods here by simply addressing that unsupervised methods can be applied to data that is relatively unstructured, or to data that has relatively few structured data categories. Note that this is an advantage with respect to the theory based methods in quadrant i and ii, which require theoretical knowledge for all analyses or data-exploration.

Potential pitfalls

As the methods in quadrant iii are mainly aimed at improving accuracy, there are research disciplines where these methods are less applicable. These disciplines are inherently in pursuit of advancing theoretical knowledge; the disciplines were quadrant iii methods stand mainly focus on engineering. The extent by which fields rely on producing 'practical knowledge' differs, and so does the applicability of methods from quadrant iii.

Quadrant iv

The core of methods in quadrant iv is that they rely on the *rule-based approach* with the goal of improving *accuracy*. These methods are aimed at good performance for specific tasks and include task-oriented / interactive systems, or specific optimizations. Examples include search-engines and automatic summarization, but also automatic approaches for anonymization. In fields that work with privacy sensitive data (such as psychotherapy and mental health), the identifiable aspects of data should be removed before data can be used for research.

Losada and Parapar (2017) used summarization to automatically select and extract psychological features based on salient sentences from psychotherapeutic texts. Trani et al. (2018) used entity linking tasks to automatically identify and link the entities mentioned in texts to their resource in an knowledge base. Schaefer et al. (2011) conducted a social-network analysis to test how extracurricular activities helped adolescents maintain existing friendships and develop new ones. Smink, Sools, Postel, et al. (2020)[2] removed all person names, locations, organizations, numbers and dates from e-mails to anonymize the data from a web-based intervention.

Strengths

Adopting a rule-based approach with the goal of advancing accuracy typically requires (advanced) expertise in computer science. On the other hand, the theoretical constructs also mandate psychological expertise on the topic. It is a great

[2]Smink, Sools, Postel, et al. (2020) is chapter 5 of this thesis.

asset to combine expertises from different fields for a multidisciplinary collaboration in pursuit of new research areas.

Potential pitfalls

However, the level of technical expertise, theoretical knowledge and interdisciplinary willingness to collaborate is one of the largest obstacles of research methods in quadrant iv. Whether or not these methods could be employed is mainly based on available technical expertise and dataset specifications. A relatively straightforward application of anonymization requires additional basic programming expertise, which usually is beyond the level of technical familiarity of psychologists.

Application

In the previous sections we discussed the analytical properties of the framework. We discuss how the framework can be applied in this section. Because the impact that the field could have on psychotherapy research is mainly dependent on the automatability of research methods, we choose the field of TCPR for this purpose. For this section, we again relied on what we think is common sense for teams working in interdisciplinary settings.

In Figure 4.6, we plotted the streams of text mining TCPR that were identified by Sools et al. (2019) in the automation-explication framework. The four streams are: *Change Analysts* (analysing emotion-abstraction patterns over time in psychodynamic therapy; stream A), *Engineers* (upscaling and technically advancing motivational interviewing research; stream B), *Explorers* (developing new and complex text mining techniques for theoretical constructs; stream C), and *Digitals* (connecting linguistic process markers to therapy outcome in internet counselling; stream D).

Stream A: Change Analysts

Stream A mainly represents the efforts of psychologists who study therapeutic change (McCarthy et al., 2017; McCarthy et al., 2014; McCarthy et al., 2011; Mergenthaler & Kachele, 1996; Pfäfflin et al., 2005; Sassaroli et al., 2014). Stream A researchers mainly study the *Therapeutic Change Model* (TCM), a software package for automatic analysis of psychotherapeutic texts (Mergenthaler, 1996). According to Mariani et al. (2013), the TCM mainly studies *referential activity* (RA), which measures "*the degree to which language is connected to non-verbal experience,*

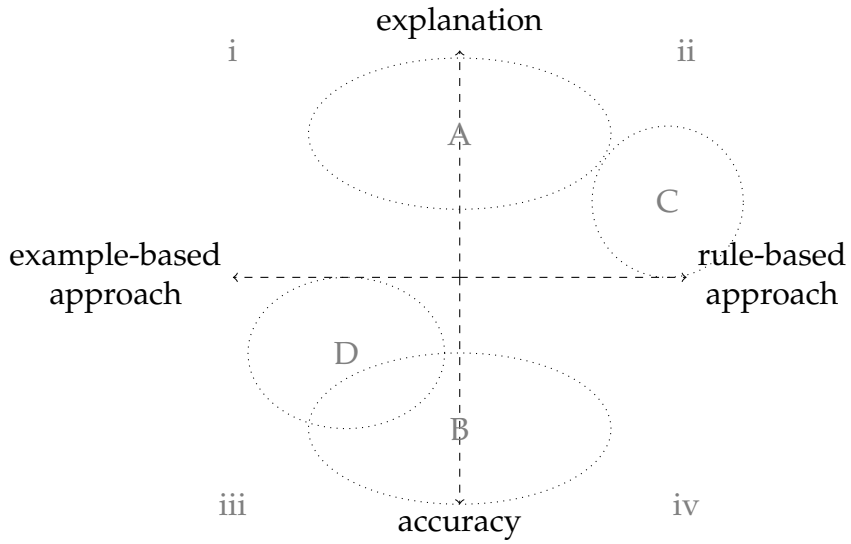


Figure 4.6: Sools et al. (2019) identified four streams of research: A) *Change Analysts*, B) *Engineers*, C) *Explorers*, and D) *Digitals*.

including bodily experience, imagery and affect" (p. 431). The four RA patterns (*relaxing*, *experiencing*, *connecting*, and *reflecting*) represent change cycles across therapy; the TCM aims to connect these cycles to treatment outcomes (Lo Verde et al., 2012).

Many researchers in stream A explore RA in the language of the client to study how intense the client is engaged with the psychotherapeutic process (Lepper & Mergenthaler, 2007). A high level of RA is characterized by a vivid language style (which evokes similar experiences in the reader or listener), whereas low RA can be identified as *vague*, *general*, or *abstract* language (Bucci & Maskit, 2006; Maskit et al., 2015). All research in stream A is primarily aimed at advancing theoretical knowledge on TCPR (and stream A relies on both approaches to automation). This is a typical example of models that are explainable.

Rule-based approaches Mergenthaler and Kachele (1996) used several "*computer-assisted measures*" to analyse past events of patients in written texts. The majority of these measures is automated through the rule-based approach. The concept *abstraction* is measured by the proportion of words with endings such as *-ity*, *-ness*, *-nce*, *-ment*, *-any*, *-ncy*, *-ship*, *-dom*, *-ing*, *-ion*, and their plural forms. Another example is *redundancy* (R), which is measured as the first-order redundancy following $R = \sum p_i \log_2(1/p_i) / \log_2(m)$, where m is the number of repeated words, and p_i the probability of the repetition. Redundancy is a clear example of the application of a rule.

Example-based approaches RA is also studied through dictionary-based approaches: the *Weighted Referential Activity Dictionary* (WRAD) is often-used (Bucci & Maskit, 2006). Stream A researchers also proposed their own dictionaries, for example, Mergenthaler (1996), used dictionaries to identify (*dis*)*pleasure*, (*dis*)*approval*, (*dis*)*attachment*, and *surprise*.

Conclusion

Because the change analysts of stream A rely on rule- *and* example-based approaches to deepen their theoretical understanding of TCPR, stream A equally belongs in quadrant i and ii, see Figure 4.6.

Stream B: Engineers

The engineers from stream B are mainly motivated by upscaling *Motivational Interviewing* (MI) research (Miller & Rollnick, 2012; Miller & Rose, 2010). Their MI corpus is partially annotated, and received additional annotations over several years with the MISC (Miller & Mount, 2001; Miller et al., 2008), MITI (Moyers et al., 2005), and YACS (K. M. Carroll et al., 2000). Many of the applications proposed in stream B require technical expertise to implement, and the majority of these researchers hold advanced degrees in computer science.

Researchers in stream B from various universities joined forces to use automated text analysis for upscaling motivational interviewing research. The commonality is that they predominantly use advanced technological applications –such as neural networks and Hidden Markov Models– to compare human and (automatic) machine annotations (Can et al., 2016; Gibson et al., 2016; Tanana et al., 2015). For example, Flemotomos et al. (2018) used deep learning to predict empathy ratings based on transcripts of the therapeutic interaction. The goal of stream B researchers often involves showing that novel technical applications works better, for example demonstrated by the title of the work of Tanana et al. (2016): "*A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing*". Many studies –such as Flemotomos et al. (2018)– in stream B have similar goals ("*we show that this method outperforms training the deep neural network end-to-end in a single stage*"), or demonstrate that a certain score is more suitable for a specific evaluation ("*our combined feature set* (which the authors proposed) *achieved a correlation of 0.56 between predicted and expert-coded empathy scores*"), see Xiao et al. (2012).

Conclusion

In stream B, technical research aims prevail (we found no substantially or theoretically guided research goals). This stream is a typical example of researchers motivated by improving accuracy rather than obtaining new theoretical insights. Because the engineers of stream B apply rule- *and* example-based approaches, they equally belong in quadrant iii and iv, see Figure 4.6.

Stream C: Explorers

Stream C is characterized by the development of new text mining approaches for complex theoretical constructs in psychology. These constructs are complex from both a theoretical and linguistic viewpoint (Cariola, 2015; Howes et al., 2012), and include the study of concepts *body boundary* construct and *repair* ("*this correlational study examines the use of words and changes in body boundary finiteness*" ..."*by measuring the strengths of associations between barrier imagery, as measured using the Body Type Dictionary*", "*we investigate whether particular types of repair can be detected from high-level dialogue features and/or lexical content, with encouraging results*"). These concepts are mainly pursued by psychologists, who –to our impression– appear to be motivated to study novel theoretical constructs (that are 'more exciting') than the currently available TCPR methods allow for.

Others stream C researchers proposed software to study theoretical constructs (Anderson et al., 1999; Bucci & Maskit, 2007; Halfon et al., 2016; Halfon et al., 2017; Nitti et al., 2010; Salvatore et al., 2012): examples include the study of *affect*, *Discourse Flow Analysis* ("*the present work is an attempt to improve quality research into the therapeutic process by means of the combination of statistical methodologies, enabling clinical interaction to be analysed*"), and *Automated Co-occurrence Analysis for Semantic Mapping* ("*the development of the efficacy and efficiency of methods of textual analysis is worth considering*" ... "*intends to contribute to such development*").

Conclusion

The explorers of stream C are motivated by developing their own alternatives to text mining, which are guided by a desire to explore theoretical constructs. Because this goal is pursued through a rule-based approach, stream C belongs in quadrant ii, see Figure 4.6.

Stream D: Digitals

Because Sools et al. (2019) identified only two studies in stream D, the following is rather tentative. A characteristic of the *born digital data* studies in this stream is that they have large sample sizes. Althoff et al. (2016) connecting linguistic process markers in the therapeutic conversation through unsupervised machine learning models. Howes et al. (2014) apply topic modelling and sentiment analysis to show that written therapy can be compared with face-to-face data.

Conclusion

Stream D relies on large example-based (machine learning) models to connect processes to outcomes in internet counselling. "Working with data" appears to be the goal, rather then obtaining more theoretically relevant features (Howes et al., 2014). Stream D therefore belongs in quadrant iii, see Figure 4.6.

Implications

It was our choice to specifically apply the automation-explication framework to the field of TCPR. The next step already presents itself: to organise a research project around the framework. For TCPR, this can be done in two ways. The first option is to extent upon the research projects that are currently conducted in the four research streams ('solidifying the existing streams'). We consider this to be the safer option, as it is inline with the standard way TCPR is being conducted now.

The more interesting possibility is to 'chart the unknown' and start research projects that lie in the unexplored (i.e. 'blank') areas of the framework. TCPR is inherently an interdisciplinary field: the framework shows that many different research disciplines are involved with the study of TCPR, but that relatively few work interdisciplinary. Let the middle of the framework represent the theoretical optimum, where all researchers agree on the research goals. Evidently, this requires an interdisciplinary approach, as the goals and methods to these goals require multidisciplinary attunement, which the framework shows is not always present in TCPR (and we expect this to be the case for more fields).

Implications for the four streams

We will now apply the two options from the previous section to the four streams. For the change analysts (stream A), the first option could imply 'business as usual', so stream A researchers would continue to invest time and effort into the

theoretical foundations that support the TCM model. The second option could imply an extension in external validity (from the client population alone) across different treatments. As stream A has a strong tradition with models aimed at explanation, another possibility would be to aim for more accurate models. As the framework shows, an increase in accuracy comes at the expense of explanation. However, as the goal of stream A is to describe, identify and analyse the essential aspects of the therapeutic process, it could be that this stream is so inherently motivated by explanation that an accuracy-based approach does not suit the research goals of stream A very well. Nevertheless, this could be an option for future research: when the TCM model is trained with a lot of data, it should be possible to use the model to predict the therapeutic process of client. This requires technical expertise, and is a step towards more interdisciplinary collaborations in stream A.

For the engineers (stream B), the first option could imply that a wider array of models is employed to further annotate the data. This would then result in more research projects that propose new routines to automatically annotate MI data, or more elegant ways to evaluate the data. The second option could imply that more effort is put into manually annotating larger datasets, so that it becomes possible to explore more theoretical goals (rather than optimizing the annotation). This requires expertise from psychologists or clinicians experienced with several of the MI coding-schemes, and is a step towards more interdisciplinary collaborations in stream B.

For the explorers (stream C), the first option could imply that the software development is advanced further. The theoretical and linguistic foundations are already laid out, so there is room for further technical developments. It is noteworthy that in stream C, the first option is quite close to the second one, as increasing the developments implies that more advanced models are used as well (now stream C appears to mainly on basic models). Doing so would automatically imply that stream C starts to use models that are aimed at accuracy, and –given that the goals now appear to be mainly of explanational nature– would imply that this is a step towards more interdisciplinary collaborations in stream C.

For the digitals (stream D), the first option could imply a further upscaling of the data samples (this stream is quite small, so this is rather tentative). The second option could imply the use of more ambiguity and creativity in the theoretical constructs (now mainly basic theoretical constructs are used). As this implies that the theoretical underpinnings of the stream are strengthened, it would mean that stream D starts to use models that are aimed at explanation, and doing so requires seeking for interdisciplinary collaborations to further advance

stream D.

The TCPR-specific goal of the framework

We hope that others will find the automation-explication framework helpful in this respect, as it helped us to give some direction to our own research projects. Based on our own experience, we known that interdisciplinary goal-setting and communication is not always straightforward to organise. In the context of TCPR: we found relatively few research groups who work in an interdisciplinary way, as the framework shows that computer scientists work in different quadrants than the psychologists (and vice versa). It is not always straightforward to organise interdisciplinary collaborations; we hope that the automation-explication framework can facilitate discussion in interdisciplinary and collaborative research projects (and gave suggestions to do so for each of the streams).

When considering automation of qualitative research methods through text mining, researchers are confronted with a choice between an example- or rule-based approach, which can be aimed at accuracy or explication. These trade-offs are described by the automation-explication framework. We showed that research disciplines differ in their preferences with respect to automation and explanation. This is also one of the reasons why interdisciplinary collaborations are difficult organise. The automation-explication framework can be helpful here as well, as it offers aspiring collaborators a thorough understanding of positions the collaborating parters. By placing the four streams of TCPR research that Sools et al. (2019) identified in the framework, we showed potential users of text mining which directions their research can take.

Conclusion with respect to text mining

From the classification of the streams it becomes apparent that there are two main contrasts: one between stream A and B, and another one between stream C and D. Stream A is placed in in the middle of both quadrant i and ii, and stream B at its polar opposite, in the middle of quadrant iii and iv. Stream C and D are also placed at each others (polar) opposites in quadrant ii and iii.

Difference between accuracy and explanation The main difference between stream A and B is the orientation of research goals: stream A consists of psychologists who use the TCM (Mergenthaler & Kachele, 1996); stream B consists computer scientists motivated by upscaling MI annotations. TCM is a text mining model based on theoretical change cycles to investigate specific aspects of

emotion-abstract language. MI is not a theoretical framework, it is a facilitation of the interpersonal relation for eliciting behaviour change, aimed at exploring and resolving ambivalence. The framework helps to map out the differences between fields. The difference between fields mainly lies in the preferences of explication (stream A relies on models that are explainable; stream B aims for accurate models), and the specific theoretical embedding.

From the classification of stream A and stream B it becomes apparent that psychologists and computer scientists pursue different goals with their data analysis. They equally rely on the rule- and example-based approach.

Difference between example- and rule-based approach to text mining Stream C and D have different preferences with respect to automation and explication. The psychologists of stream C appear to be motivated by developing new, theoretically and linguistically complex text mining approaches. These explorers values some theoretical construct to such a great extent, that they developed their own alternative to text mining. Yet, stream C mainly relies on straightforward applications of rule-based approaches to operationalize (or propose) new constructs.

The computer scientists of stream D are mainly dedicated to investigate data from chat- and SMS-texting-based counselling. The central theme of the research of these digitals appears to apply text mining to data that is already available and collected. Their goal is to finding the best (or optimal) text features, rather than obtaining more theoretically relevant features (i.e. "*finer-grained lexical features*"). The goal of stream D is technical, and consists of finding the best ways to model the data through the example-based approach, with differences with respect to accuracy and explication.

From the classification of stream C and stream D it (again) becomes apparent that psychologists and computer scientists pursue different goals with their data analysis. Perhaps, psychologists mainly use rule-based approaches because they are more straightforward to implement, psychologists mainly use rule-based approaches. More advanced example-based models are predominantly used in stream D.

General implications of the framework

When considering a text mining alternative to a qualitative research method, the automation-explication framework describes two trade-offs: explication and automation, with four (crossed) variations. The most accurate predictions are typically achieved through complex 'black-box' models. These are often crit-

icised because their inner workings are non-explainable. Explainable models on the other hand are easy to understand, but impoverished in their predictive accuracy.

Combining this trade-off with rule- and example-based approaches aids researchers in deciding whether their methods (and data) are suitable for automation. It is difficult to automate qualitative research methods through a text mining approach because there is a difference between the goals and models of the *developers* of (new) text mining applications (who are mainly computer scientists), and the *users* of these text mining applications (such as psychologists). In the current paper, we identify two trade-offs that can give insight in the differences between these preferences: the orientation of *explication*, and the method of *automation*. We introduce the *automation-explication* framework, which can bring users and developers text mining methods together and can stimulate interdisciplinary discussions. As a specific use-case, we use TCPR as an example.

Although we did not directly address the difference between qualitative and quantitative methods, automated text-analysis methods could bridge these branches of research, or at least form the basis for an interdisciplinary discussion. Texts –or other forms of verbatim texts– are usually considered to be the domain of qualitative research, but large bodies of text mandate the use of quantitative methods. Depending on the application or research question, the model should have the ability to explain why certain predictions are made. Qualitative research methods almost always offer this insight, but (highly accurate) text mining models do not always have this property. It is not very useful to know which individuals changed over the course of therapy without any explication of what exactly changed. So what should be used for text mining? The automation-explication framework describes that this is typically a trade-off.

Strengths and limitations

The different orientations of explication are well-known throughout the literature. However, for automated approaches to qualitative research, this trade-off appears to be less well-known although it –arguably– is even more relevant. Text mining essentially involves the analysis of language, which draws researchers from many disciplines, such as the humanities, linguistics, statistics, and computer science to name a few. Understanding the trade-off between automation and explication can help researchers from different disciplines understand the strengths and limitations of each others preferences, as it is common that funding for scientific grants requires on interdisciplinary collaboration. The framework demonstrates that that it is not as simple as it may seem, as researchers

from different disciplines have their own applications, and it is good to create more awareness about the difficulties that surround interdisciplinarity.

The framework is based on the assumption that automation and explication are two different and independent concepts. However, the way the four streams are classified could indicate that explication and automation are not as independent as we assumed. The framework indicates that the example-based approach predominates accuracy orientated research, whereas the rule-based approach frequent in explainable research. The technological advancements of text mining appear to be in line with this observation: throughout the history of machine learning (the *statistical revolution*) example-based models became more prominently featured in text mining than rule-based models (Johnson, 2009; Jurafsky & Martin, 2014; Manning & Schütze, 1999; Martinez & Martinez, 2015).

So, are automation and explication independent constructs? We have to conclude that –even though we still think that from a theoretical point of view these constructs are independent of each other– perhaps this distinction is more difficult to maintain in practice. The field of automation was –around the 1990's– in a transition away from rule-based methods to example-based methods because these models worked better in practice. This is a limitation of our framework: what we presented to be two perpendicular concepts are perhaps oblique in practice.

Nevertheless, even though we cannot present empirical evidence for this claim here, we do feel that automation and explication are two independent constructs. The difficulty here is that more often than not, the availability of (annotated text-)data determines the direction of the research projects (Smink, Sools, van der Zwaan, et al., 2019). So, the framework describes the ideal situation at the beginning of a text mining project, before data is collected. Then, researchers are confronted with two choices: what kind of model explication is desirable, and through which automation approach should this be operationalized. How automation and explication relate to each other is determined by the data that is used in practice.

We feel that this specific point deserves more attention, and that this is the greatest strength of our framework. If it is true that the example-based approach is more often associated with accurate models, than perhaps then a *machine learning / data science / Big data* approach sounds good on a grant application, but it could fundamentally not be what the research project requires if theoretical goals are being pursued. We hope that the framework will contribute to this line of thinking and advocates a nuanced understanding of how the practice of Big data interacts and changes the psychological science.

Transferability

It was our choice to focus on TCPR as a specific use-case, which means that some of our conclusions that we draw are specific to TCPR, and are less *generalizable* to other research disciplines. However, we base our TCPR findings on reviews of the literature, so we feel confident that we indeed covered the majority of the relevant TCPR field. Although some of our conclusions are perhaps not so generalizable, this –in no way– implies that the our proposed framework is not *transferable*. Both approaches to automation and the explication are deeply embedded in the literature, which establishes that these discussions are also relevant elsewhere (that is, outside of TCPR). If other fields match some of the key-traits of TCPR that we shall discuss here, it should be straightforward to apply the framework in another field (Smaling, 2003).

The first aspect is that the other research discipline should predominantly rely on text-analysis as well. In other words, the field should be predominated by qualitative research methods. Additionally, automation should be a relevant theme, so there should be attempts (or at least the intention) to automate some of the available methods. So, for transferability, it is important that the research discipline is in some sort of similar transition away from qualitative methods towards more automated methods. To apply the framework in a similar way to what we did, there also needs to be a some sort of literature review available. We based our work on a scoping / state-of-the-art review, but we expect more types of reviews are suitable.

Future possibilities

Because so many fields could potentially benefit from text mining, it should come as no surprise that the ideal way to cover the field is by describing the field in terms of automation and explication. Although we are confident that these two discussions indeed span the important topics, it is of course possible to plot alternative trade-offs. The characteristics of a concept that could be used in a similar framework are that it: a) in fact, has a trade-off consisting of two alternatives, which b) can be organised as two opposing ends of a spectrum, with c) hybrid possibilities in between the two ends. A few examples of other (potentially interesting) trade-offs are: the difference between *supervised* and *unsupervised* research methods, the distinction between *theory-* and *data*-driven methods, and the variation between (statistical) conclusions adhering to either the *group* or the level of the *individual*.

Part II

TCPR models

What
Works
When
for
Whom

The term [Therapeutic] Change Process Research ([T]CPR) was introduced more than 20 years ago to refer to research that overcomes the old process-outcome dichotomy by focusing "***on identifying, describing, explaining, and predicting the effects of the processes that bring about therapeutic change***" (Greenberg, 1986, p.4).

Elliott (2010)

What
Works
When
for
Whom

5 Analysis of the e-mails from the *Dutch* Web-Based Intervention '*Alcohol de Baas*': Assessment of Early Indications of Drop-Out in an Online Alcohol Intervention

Abstract

Nowadays, traditional forms of psychotherapy are increasingly complemented by online interactions between client and counsellor. In (some) web-based psychotherapeutic interventions, meetings are exclusively online through asynchronous messages. As the active ingredients of therapy are included in the exchange of several e-mails, this verbal exchange contains a wealth of information about the psychotherapeutic change process. Unfortunately, drop-out related issues are exacerbated online. We employed several machine learning models to find (early) signs of drop-out in the e-mail data from the 'Alcohol de Baas' intervention by Tactus. Our analyses indicate that the e-mail texts contain information about drop-out, but as drop-out is a multidimensional construct, it remains a complex task to accurately predict who will drop-out. Nevertheless, by taking this approach, we present insight in the possibilities of working with e-mail data and present some preliminary findings (which stress the importance of a good working alliance between client and counsellor, distinguish between formal and informal language, and highlight the importance of Tactus' internet forum).

Keywords: Therapeutic Change Process Research (TCPR), alcohol use disorder (AUD), drop-out, web-based psychotherapeutic interventions, e-mail data

Smink, W. A. C., Sools, A. M., Postel, M. G., Tjong Kim Sang, E., Elfrink, A., Libbertz-Mohr, L. B., Veldkamp, B. P., & Westerhof, G. J. (2020). *Under review*.

Introduction

As addictive dependencies have a global impact (Whiteford et al., 2013), almost everyone will have some degree of experience with an addiction, with alcohol use disorder (AUD) as the prevailing substance abuse disorder (Degenhardt et al., 2018). It is estimated that –around the world– 283 million individuals suffer from AUD, representing around 5.1% of all adults (Hammer et al., 2018). As these numbers are predicted to increase globally –especially in low-income countries (Whiteford et al., 2013)– the need for accessible therapy for AUD is now more apparent than ever. Web-based psychotherapeutic interventions have established themselves as a valid intervention for AUD (Pedersen et al., 2016; Postel, de Haan, & de Jong, 2010), it is –however– well-known that these online alternatives to treatment are plagued by high rate of drop-out (Kelders et al., 2012; Melville et al., 2010). The problems of high drop-out in drug- and alcohol-treatment are frequently reported (Copeland & Hall, 1992; Schroder et al., 2009), and these problems are exacerbated online (Kelders et al., 2012). It is our aim to see whether the e-mails from a web-based psychotherapeutic intervention –or 'online psychotherapy', a variety of different terms address what is essentially the same concept (Lau et al., 2013; Oh et al., 2005)– are helpful to obtain a better understanding of drop-out.

Our specific interest in web-based interventions for AUD comes from the possibility to target the (potentially AUD) affected early on. There is large delay between the onset of alcohol dependency and first treatment contact (Chapman et al., 2015). Even when there is a high-density and high accessibility of primary and specialised care, the treatment delay is –on average– up to 18 years (Bruffaerts et al., 2007; Chapman et al., 2015; Korbmacher, 2014). With only one in three problematic drinkers ever seeking treatment (Cunningham & Breslin, 2004), it is safe to say that those who are alcohol dependent keep to themselves: more than one-third of those who feel they have problematic drinking behaviour had a medical or mental health visit, but more often than not their drinking behaviour was not discussed (Weisner & Matzger, 2003). However, a growing number of those who expect they have some form of AUD resort to online solutions for their drinking problems early on (Cloud & Peacock, 2001; James et al., 2018). It is therefore not strange that traditional face-to-face therapy is in transition towards (more) online supplements, as the affected are more active online nowadays (Vernon, 2010).

Web-based interventions for AUD

Some *general* advantages of web-based interventions are that they have a lower threshold for first treatment contact (Amichai-Hamburger et al., 2014; Hoogendoorn et al., 2017), some online interventions are as effective as traditional face-to-face therapy (Barak et al., 2008; Gainsbury & Blaszczynski, 2011; Howes et al., 2012), come at low-cost (Schweitzer & Synowiec, 2012), and have usually no to (relatively) short waiting lists (Amichai-Hamburger et al., 2014). Online many clients feel they can maintain their privacy (Berger et al., 2005), feel less stigmatized (Postel, de Haan, ter Huurne, Becker, et al., 2010; Rooke et al., 2010), and (sometimes even) prefer the impersonal nature of the web, as they do not have to disclose their feelings and problems in person (Griffiths, 2005). Online interventions for substance abuse form a large part of the online offer, with many targeting AUD specifically (M. A. A. Rogers et al., 2017).

The *specific* advantage of web-based interventions for AUD is the all-time availability. Websites can be accessed every hour of the day (and every day of the year), which is a great advantage over face-to-face treatment for those who cannot attend to treatment at business hours (Moritz et al., 2013). This is of special importance when treating AUD (Cloud & Peacock, 2001), as the willingness of a client to change his (or her) drinking behaviour is often of volatile nature and easily affected by (negative) events, that can –for example– also occur during holidays (Cunningham et al., 1994). Even though web-based interventions make it difficult for a counsellor to react to the non-verbal cues of clients, they are a helpful and welcome addition to the treatment of AUD.

There is no debate about the biggest drawback of web-based interventions (Fernández-Álvarez et al., 2017; Karyotaki et al., 2015): all online interventions are plagued by a high rate of drop-out, which is on average 50% (Kelders et al., 2012). Online treatment adherence is sometimes as low as 1% (Christensen et al., 2009; Farvolden et al., 2005), with drop-out rates going up to 99% (Andrade et al., 2016; Karyotaki et al., 2015). The problem of high drop-out for the online alternatives for AUD specifically are well-known as well (Copeland & Hall, 1992; Kelders et al., 2012; Schroder et al., 2009), and frequently addressed in the literature (Copeland & Martin, 2004). To give an impression of the online drop-out rates for AUD: Postel (2011) reported a drop-out rate of 54%, and 84.5% dropped out in the study of Linke et al. (2007).

Drop-out

We did not find studies that set out to specifically address the reasons for dropping-out of an online intervention for AUD (if reported, it usually appears

to be one of many complementary analyses). Several qualitative investigations report that clients drop-out because they have no reliable computer access, lack computing experience, face computer or Internet problems, dislike computers, prefer face-to-face treatment, or lack the necessary peace and quiet in their computing environment (Stark, 1992). Drop-out is referred to in the literature as *premature termination*, *non-usage*, *low attrition* or *retention* (Lau et al., 2013; Melville et al., 2010; Oh et al., 2005). Who should be labelled as a drop-out is not universally agreed upon: some argue that only those who did not finish the complete intervention are drop-out, others argue that drop-out are all the clients who did not reach a certain cap of required attended sessions (Swift & Greenberg, 2012), and some say that only the judgement of the counsellor can determine who is a drop-out, as it is possible that a client dropped out because he or she already experienced the beneficial effects of therapy (Krishnamurthy et al., 2015; Zandberg et al., 2016).

These distinctions matter, as the requirements for treatment completion affect the drop-out rate (Yeung et al., 2015). In line with Eysenbach's *Law of Attrition*, in this study, the drop-out are those who did not finish all the treatment sessions that required to complete the treatment protocol (Eysenbach, 2005). Aside from difficulties in definition, the studies that analyse drop-out often use different sample groups, diffuse treatments, and diverse subtypes of disorders (Khazaie et al., 2016), which makes it even more difficult to compare drop-out between studies. This is a shortcoming in the (AUD-)literature: knowing who is likely to benefit from the intervention provides a better basis for an evidence-based allocation of clients to treatment (Lincoln et al., 2014). We are not aware of studies that rely on text mining (or other automated) approaches to study drop-out.

'Alcohol de Baas'

We will use the web-based intervention '*Alcohol de Baas*' ('AdB'; translated from *Dutch* as 'Look at your drinking') as the use-case example (Postel et al., 2008; Postel, de Haan, ter Huurne, Becker, et al., 2010). The goal of AdB is to reduce the alcohol consumption of clients who registered themselves with (self-reported) drinking problems. AdB is an evidence-based intervention for treating AUD: Postel (2011) demonstrated that the intervention led to a significant decrease in alcohol consumption, that was maintained at a six months follow-up (p. 14, 66−67, 74, and 95). According to Postel (2011), AdB had 897 users in 2005, and 885 users in 2009 (p. 68, 118; which we consider to be a substantial interest given the size of the Dutch population). AdB is rooted in cognitive behavioural

therapy and motivational interviewing (p. 14), both empirically substantiated approaches for the treatment of substance abuse (Miller & Rollnick, 2012; Miller & Rose, 2010).

The structure of web-based interventions often adheres to client and counsellor exchanging e-mails (Chester & Glass, 2006; Rochlen et al., 2004), and AdB is no different in this respect. Yet, relatively few studies actually analyse the content of the e-mails (Smink, Sools, van der Zwaan, et al., 2019; Sools et al., 2019):[1] more often than not, the difference between (several) pre- and post-therapeutic measurements is studied (Smink, Fox, et al., 2019).[1] This creates a gap between the information that is available, and the data that have been used. The content of the e-mails potentially contains a wealth of information about several open questions and issues, for example about drop-out (which is also substantial in AdB). The AdB intervention has one important distinction with other web-based interventions: it was free, meaning there were no financial reasons for dropping out. Therefore, Postel, de Haan, ter Huurne, Becker, et al. (2010) concluded that the main reasons for drop-out of AdB were personal and unrelated to the program. To understand drop-out better, we resort to the early e-mails of the intervention, to see if these texts can be helpful to discriminate between drop-out and completers.

Studying the therapeutic exchange

So, why and how can drop-out be studied by looking at e-mails? Throughout the history of psychology, language has established its importance for understanding the human mind (Greenberg, 1986; Hill & Lambert, 2004), and it is no surprise that the verbal exchange between client and counsellor is the cornerstone of almost every 'on'- and 'offline' psychotherapeutic intervention (Anderson et al., 1999; Muntigl & Horvath, 2005). Language gives the unique opportunity to express and communicate emotions, thoughts, motivations, and intentions (Tausczik & Pennebaker, 2010). It is well-known that constructing a life story has therapeutic effects, as it can help to process disturbing life events and result in less rumination (Pennebaker & Seagal, 1999). Also, the mere act of articulating emotions and troublesome thoughts is beneficial to (short-term) improvements of mental well-being (Bucci, 2013). This result has been replicated for AUD: a verbal reflection on alcohol consumption can noticeably reduce the consumed amounts (Kypri et al., 2007; Meier et al., 2018; Schrimsher & Filtz, 2011).

Therapeutic Change Process Research (TCPR) is concerned with studying

[1] Smink, Sools, van der Zwaan, et al. (2019), Sools et al. (2019), and Smink, Fox, et al. (2019) are chapter 2, 3, and 6 of this thesis.

the 'active ingredients' that underlie these (change) effects (Greenberg, 1986), as the field aims to "*identify, describe, explain, and predict the effects of the processes that bring about therapeutic change over the entire course of therapy*" (p. 4). An application of TCPR for AUD is the work of Blankers (2011), who compared the drinking behaviour of individuals following an online self-help course with either automatic or counsellor feedback. The goal was to improve self-control, management of cravings, and (realistic) goal setting. In the three months follow-up, Blankers (2011) found that the feedback of the counsellor could be associated with the strongest drop in alcohol consumption (Postel, de Haan, ter Huurne, Becker, et al., 2010). We are not aware of many other applications of TCPR for AUD, as [T]CPR and related terms are not widely used as of yet, making it difficult to identify and find these studies (Elliott, 2010; Greenberg, 1986; Smink, Sools, van der Zwaan, et al., 2019).

Because –in the case of AdB– the client and counsellor primarily used e-mail to establish the beneficial effects of e-mails, the therapeutic process should be included in these e-mails. The exchanged verbatim is therefore a valuable source for studying the beneficial therapeutic change processes in clients. However, as the drop-out issues that plague traditional interventions are exacerbated online, we study the AdB e-mails to see if they contain any early 'warning' signs of drop-out. Because drop-out usually is so high, we focus on the early e-mails from the intervention (i.e. the first four), because these are generally available for the majority of the respondents. To the best of our knowledge, this has not been done before and we did not find applications of text mining applications specifically tailored to study drop-out for AUD.

Text mining TCPR

In our view, text mining can be defined as a broad methodological framework (rather than just one specific research method) that entail three types of tasks (Hoogendoorn et al., 2017): counting the frequency of words in texts, counting words into pre-defined categories ("*coupling words to domain-specific ontology*", i.e. dictionary-based approaches, the *Linguistic Inquiry and Word Count* –LIWC– program is a prime example here), and identifying the topics in a text. As all ingredients that elicit therapeutic change within clients should be within the written correspondence, we are curious to see what we can learn more about drop-out through some basic text mining approaches. We will therefore do two things: first, we will 'simply' count the most frequently used words in e-mails through unigram models (Lioma & Keith van Rijsbergen, 2008), and second, given the popularity of dictionary-based approaches in psychology (Sools et al.,

2019), we also use LIWC.

The dictionary-based program LIWC by Pennebaker, Boyd, et al. (2015) has perhaps the most widespread use in (general) psychological research. Aside from the fact that LIWC is easy to use, it is also available in many other languages than English alone. As LIWC is available with demonstrated validity in Dutch (Boot et al., 2017; van Wissen & Boot, 2017), we choose to use LIWC. LIWC encompasses many aspects of the e-mail texts (such as the function words and punctuation) and calculates the degree to which various categories of words frequent in a text, which allows for the exploration of emotional, cognitive, and structural components of verbatim (Pennebaker et al., 2003). See Liehr et al. (2010) for an example: they assessed self-change by applying LIWC to written stories and studied stressful feelings over the course of an intervention for substance abuse.

Research goal

Drop-out plagues the evaluation of all types of treatments, but especially AUD interventions. First treatment contact is (on average) years later than the onset of problems, but online help is sought earlier. Unfortunately, the problems of drop-out are exacerbated online, so there is clear value in understanding drop-out in web-based interventions. TCPR is a research discipline that is concerned with understanding the underlying processes that could affect drop-out, and text mining methods provide means to analyse large text corpora. We will (re-)analyse the data from AdB, an online intervention aimed at reducing the alcohol consumption of clients who registered themselves with (self-reported) problems. Our goal is to find which textual aspects in the early e-mails improve our understanding of drop-out.

Method

Aside from discussing our own study, we have to address the original *Alcohol de Baas* (AdB, loosely translated from *Dutch* as 'Look at your drinking') study here as well (Postel, 2011). We will first describe the clients we included, see Figure 5.1 for the flow of drop-out. For a more in-depth characterization of the clients, we refer the interested reader to the four case descriptions that Giesler (2019) and Krstić (2019) provided in the Appendix.

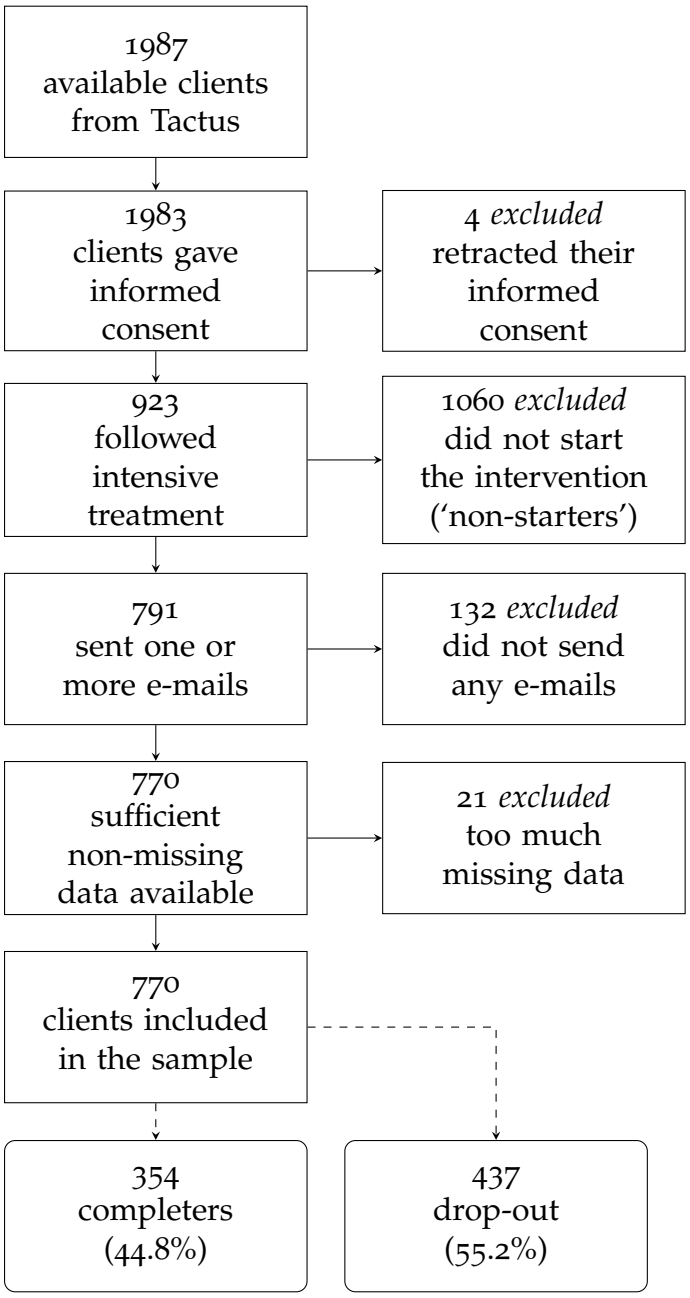


Figure 5.1: Flowchart of the (excluded) clients.

Respondents

Tactus used broad criteria for the inclusion of clients: everyone over 16 (which was the legal drinking age in the Netherlands) could register and participate in AdB. Prior to starting the treatment, all respondents gave their informed consent that their data could be (re-)used, but had the right to withdraw at any moment (which four respondents did, see Figure 5.1). Postel, de Haan, ter Huurne, Becker, et al. (2010) received ethical approval for (re-)analysis of AdB. Not all of the available 1987 clients were eligible for inclusion, we had to exclude

Table 5.1: Overview of the client's age, years of problematic alcohol consumption, and the average units of consumed alcohol.

	Drop-out				Completer			
	M	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.
Age	44.4	11.0	17	78	47.7	10.2	19	75
Cons. years.	17.9	10.5	3	35	19.3	11.5	5	40
Av. cons. alc.	8.1	7.0	0	25	7.0	5.0	0	24

several clients (see Figure 5.1). In total, 770 clients were included in our study, their median age was 46, with age ranging between 17 and 78 years old, see Table 5.1. See Table 5.2 for an overview of the demographic characteristics of the clients.

The majority of the clients we included reported to identify as female, were of the Dutch nationality (and spoke Dutch), were married, and finished a higher vocational degree. They smoked occasionally, but did not use drugs, nor did they gamble. Their main reason to start with treatment was because they worried about their drinking behaviour: the median consumed units of alcohol pear week was 36 (at the beginning of the program). About half of the sample frequently experienced feelings of depression or other psychological problems (Cunningham et al., 1994). For an overview of the physical ailments of the respondents, see Table 5.3.

Procedure

We present an overview of the procedure in Figure 5.2. The data available to us included the clients up to 2017, who registered themselves for AdB with self-reported alcohol problems (Postel et al., 2005). The data contained the responses of the clients on several questionnaires (administered 'pre' and 'post' of the intervention), and the e-mails that the clients and their counsellors exchanged. Depending on the part of the program, the e-mails are more (or less) tailored to each individual client. AdB consists out of two parts: the first focuses on drinking habits, the second on behaviour change (Postel, de Haan, ter Huurne, Becker, et al., 2010). The goal of AdB is to reduce drinking, or stop alcohol intake completely (Postel, de Haan, & de Jong, 2010).

In the first part of AdB, the counsellor helps the client to analyse his (or her) drinking habits. This is done through several assignments and assessments, that

Table 5.2: Overview demographic characteristics in-take questionnaire, split to drop-out and completer.

Variables	Drop-out ($N = 424$)		Completer ($N = 346$)	
	N	%	N	%
Gender				
males	209	61.3	133	38.7
females	215	50.2	213	49.8
Nationality				
Dutch	22	44.9	27	55.1
no answer	402	55.8	319	44.2
Education				
primary	5	55.6	4	44.4
lower vocational education	0	0.0	0	0.0
school of higher general secondary education / pre-university education	56	62.9	33	37.1
intermediate vocational education	103	60.2	68	39.8
higher vocational education	137	52.5	124	47.5
university	40	40.8	58	59.2
no answer	11	0.1	19	0.0
Ever followed treatment before				
yes	25	48.1	27	51.9
no	308	56.2	240	43.8
no answer	91	53.5	79	46.4
Reason starting the intervention				
I think I am drinking too much	334	59.6	264	40.4
I want advice about my alcohol consumption	14	58.3	10	41.7
something happened and I want to change my drinking behaviour	36	53.7	31	46.3
others think I am drinking too much	14	70.0	6	30.0
other reasons	24	45.3	29	54.7
no answer	2	25.0	6	75.0
Tobacco use				
never	164	46.7	187	53.3
now and then	36	60.0	24	40.0
daily	224	62.4	135	37.6

Table 5.3: Overview physical ailments assessment-questionnaire, split to drop-out and completer

Variables	Drop-out ($N = 424$)		Completer ($N = 346$)	
	N	%	N	%
Diarrhoea				
yes	286	53.9	245	46.1
no	138	57.7	101	42.3
Epileptic seizures				
yes	14	70.0	6	30.0
no	410	54.7	340	45.3
Memory problems				
yes	322	53.4	281	46.6
no	102	61.1	65	38.9
Palpitations				
yes	218	52.8	195	47.2
no	206	57.7	151	42.3
Headache				
yes	277	53.4	242	46.6
no	147	58.6	104	41.4
Hyperventilation				
yes	97	59.9	65	40.1
no	327	53.8	281	46.2
Stomach ache				
yes	223	54.9	183	45.1
no	201	55.2	163	44.8
Fatigue, lack of energy				
yes	374	55.4	301	44.6
no	50	52.6	45	47.4
Sexual problems				
yes	193	51.2	184	48.8
no	231	58.8	162	41.2

are followed-up by feedback from the counsellor. The first part ends with a personalized advice to the client. The second part of the intervention focuses on changing the dysfunctional drinking habits of a client, and aims to replace the thoughts associated with alcohol cravings by more helpful ones. After 10 weeks (but the duration varies slightly between clients), AdB ends with the formulation of an action plan for maintaining the new drinking behaviour or sobriety to prevent relapse. Table 5.4 contains an outline of the treatment protocol with the activities for each week, accompanied by quotes from the the e-mails between client and counsellor.

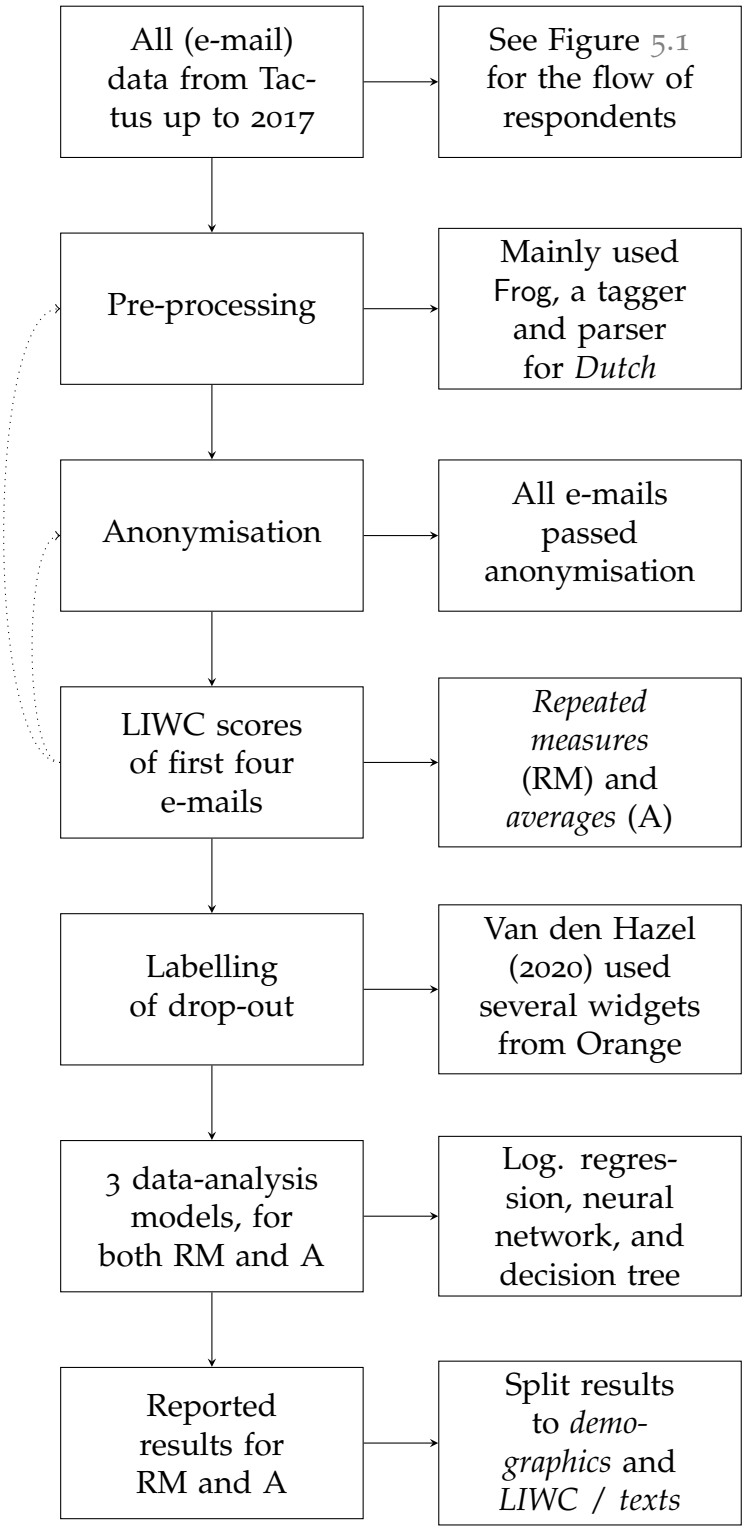


Figure 5.2: Overview of the procedure. The pre-processing and anonymisation were adjusted and optimized several times.

Table 5.4: Overview of the content of the intervention with typical quotes from each week of treatment. Table comes from Libbertz-Mohr (2020).

Treatment Procedure	*Content*	*Counsellor*	*Client*
Beginning	Welcoming	Dear PER, welcome. My name is PER and I am counsellor for LOC. In this message I will talk about your registration and discuss the program in more detail (...).	Dear PER, I checked this only until this morning. I still have to adapt to how this is working. There is something wrong with my computer. It says you sent me another message earlier and some homework assignments, so I am looking for it right now.
Part I	1. Advantages and disadvantages of drinking	In the list of medical conditions, you mention that you sometimes feel overly tired. Did you already talk with your general practitioner about that? This problem can be related to alcohol consumption. Less alcohol gives you more energy. What do you think about that?	Alcohol helps me tremendously to relax after a stressful day. Even stronger, I find it difficult to relax without alcohol. (...). Distorted sleeping rhythm, if a bottle of wine or beer is on the table, I am no longer willing to go to bed. I keep on drinking until I am dead tired.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	2. When and how much do you drink?	I see that your consumption is equally high in weeks that you work and in weeks that you stay at home and your consumption was lower when you were on vacations. Just as you had predicted.	My girlfriend is anti-alcohol and anti-smoking. I do not smoke but every now and then I like having a beer. Now, during the weekends when we are together, I do not drink. However, during the week I am alone. (...). When I come home and I do not have to work the following day or have a nightshift, I grab a beer. Yet, that beer becomes easily NUM and NUM beers.
	3. Analysis of drinking situations	That was an awful situation. You said that deep inside you something began to tremble. In your moment description I read that you drank a glass of wine to calm down. Did the glass of wine have the effect you had anticipated? What did you do after you drank the glass of wine? What did this do with your feelings?	My stumbling block is that I do not feel well when I come home from work. The work itself is fun, but my director is just an idiot. When I come home stressed and do not feel well, I will drink some glasses of wine. It is like a break for me, just like NUM years ago when I was still smoking.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	4. Recognize drinking situations	Based on the different situations that you described it seems that you often feel like having a drink when you are: out of house, at the golf club, at home with your partner, in a restaurant or café, at a party, at weekends. These we call your risk-situations. In these situations, you seem to have a higher urge to drink.	I had a partner that liked to drink. Now, there are no days on which I do not drink. Currently I drink some more because the vacation period is over and there are other circumstances like some nice parties and gatherings. I almost always drink NUM or NUM beers after work. Additionally, I always have a glass of red wine when eating. Mainly around NUM o'clock or between Num and NUM in the evening. What I drink above of that is mainly due to being in social gatherings.
Part II	5. Set a drinking goal	The choice to drink less is a good choice. I think that this is achievable (...). You can always drink a little. This means that you don't have to get nervous when you are offered a drink somewhere. Also, you avoid questions by others that you might consider uncomfortable.	I do not want to reach the point where a doctor tells me that due to a cirrhosis of the liver I may never drink again. (...). In short, I do not want to set myself up for a severe alcohol addiction and detox clinics. I want to try to drink on a 'not harmful to health' level for half a year.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	6. Identifying unhelpful & helpful thoughts	You are going to investigate your own unhelpful risky thoughts and try to change them into more positive helpful thoughts. Helpful thoughts are thoughts that aid you in achieving your goals and that support you in your plans. Unhelpful thoughts tempt you to drink more than you had intended. (...). One risky thought you mentioned earlier was 'Now I really need wine to get me through this'	The function of alcohol after parties, or other events, was like a reward. Or if I had an annoying meeting then I had at least something to look forward to at home. (...). You asked whether I wanted to think of helpful thoughts.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	7. Formulate helpful behaviors for moments of craving	Just like your thoughts, behaviors can also be distinguished in helpful and unhelpful behaviors. An unhelpful behavior can be that you accepted the glass of wine that was offered to you at a family dinner. Such a risk behavior does not help you in moments you actually do not want to drink. It is better to do other things instead. Searching a distraction or doing something nice can help you overcome your craving. It can be difficult to come up with alternatives just when the craving arises, so it can be useful to think of some activities beforehand and make a list of them. You do not have to choose complicated activities and even simple ones like vacuum-cleaning or taking a shower can be very effective.	I bought a training DVD and restored my stepping-machine. When I feel stressed, I will use the stepping-machine for NUM minutes and for NUM days a week I train with the DVD.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	8. Decision-making	For the task 'decisions-making' you create a schema of Decision-making moments that you go through before you decide to drink. For each moment in your decision-making there is the chance to step out. Read carefully and try to imagine what you can or could do at the different moments of decision-making.	NUM days ago, I wanted to drink a glass of wine. It was after lunch and I really felt that I needed to sit down with a glass of wine to relax. Yet, I felt that I had made so much progress in the last weeks and I did not want to risk it. I decided to try and make a cup of coffee instead and drink it. It helped and I did not need to have the wine.
	9. Formulating an action plan*	We now reached step NUM, creating an action plan, so that you have something that can serve as a guide XXXX and to prevent a relapse. You have read that there are moments when you can take other actions instead of choosing to drink. For the task 'actionplan' you will make a short, condensed overview of your personal motivation as well as your helpful thoughts and behaviors. It is good to have this all together in one place.	I am curious how we will proceed with the action plan and how this can increase my efforts to reach my goal. I am also interested in where you see my personal pitfalls and what, according to you, the mechanisms are that make me drink more than I plan.

Table 5.4 continued from previous page

Treatment Procedure	*Content*	*Counsellor*	*Client*
	10. Wrapping up	Dear PER, thank you for the message. I like reading that the moments in which you want to drink but do not become more and more normal to you. Also, that you perceive the advantages of not drinking as greater to the advantages of drinking. We will wrap up the treatment here and therefore I want to give you a summary of the treatment. (...). For the next half year, you can always come back online to check your files and to re-read what we did. This can help you in solidifying the changes that you achieved. (...). For now, I wish you much success and all the best!	Dear PER, (...). Thank you for the conversations we had the past weeks. The fun thing is, that what you wrote in 'Subject' (Subject of the e-mails) always described the essence well. Especially in your last mail. 'Taking the matters in your own hand'. You are exactly right, and I think that this is the process that I have started. Thank you, I am on a good path and I think it will become better and better. I want to keep on improving. Thank you again!
Aftercare & Conclusion		Dear PER, it is nice to read that you are feeling well, and things have turned out just as you had hoped for.	I am feeling much better than I had expected. The life I had seems to be so far away now, instead of the actual NUM months. I never want to go back.

Note. Names, dates, numbers, locations, medical problems, and other ('miscellaneous') entities are replaced by the abbreviations 'PER', 'DATE', 'NUM', 'PRO', and 'MISC' respectively.

Pre-processing

Before we could conduct our analyses, we had to pre-process the e-mails to restructure them into a format suitable for data-analysis. We mainly used *Frog* (van den Bosch et al., 2007) and NLTK (Bird et al., 2009) to do so. NLTK is a Python library that has detailed documentation (Bird et al., 2009), which is easily accessible for those with little to no programming experience. Python itself is an interpreted, high-level, and general-purpose programming language (Python Software Foundation, 2020). NLTK tallies sentences by counting word-terminal end-of-sentence punctuation like the period, question- and exclamation marks.

NLTK has a list of abbreviations, which are not included in the punctuation and sentence count. Word-internal punctuation, like the first period in "e.g.," is ignored. Handling of interjections depends on their punctuation, for example, "Oh?" is a separate sentence while "Oh," is part of the following sentence. Sentence fragments and quotes with end-of-sentence punctuation are counted as separate sentences.

We first divided the text into tokens and sentences. Some clients included texts from a previous e-mail by the counsellor in their own e-mails. As these quotes were not written by the client, we removed these from earlier e-mails (making use of Python). We only considered texts containing twenty words or more, and then normalized all texts by converting all capitals to lower case characters.

Anonymisation

As can be seen in Table 5.4, the names, dates, numbers, locations, medical problems, and other ('miscellaneous') entities were replaced by the abbreviations 'PER', 'DATE', 'NUM', 'PRO', and 'MISC' respectively (Tjong Kim Sang et al., 2019). As the identity of the client could be revealed though these entities, we used the *Frog*-program for *named entity recognition* and for anonymization of these entities (van den Bosch et al., 2007). All data we used passed the anonymization process, see Figure 5.2. For example, the sentence "*Mary went home earlier for Bob and for John, but also for Lisa, because she had her second day of exams*", was changed to "PER *went home earlier for* PER *and for* PER, *but also for* PER". Below, we give an example of an anonymized (and translated) e-mail (Van den Hazel, 2020):

> "Dear [PER], my name is [PER]. I am going to try my best to answer your questions, though there are quite many! I think the quantity has stayed the same these last years. Of course, I have looked on

> this website before, and I know I drink too much. I think that it has fluctuated in the past from some to a lot of glasses per week. So I actually tried to monitor it for a while, which is is why the quantity is stable (though still too high). For the onset of my drinking habits I need to go back to my past. I am from a traditional family where smoking and drinking were not allowed. So I did not drink much in my childhood. From a young age, I socialized in groups where (heavy) drinking was the norm."

As can be seen from the example, the anonymisation removed personal information, while it remained possible to understand what the e-mail was about. Because named entity recognition is a machine learning task, the anonymisation procedure was not without flaws. In part, this was due to the fact that entities could be misspelled (for example, the city '*Asmterdam*' is not recognized as a location). To ensure that all personal information was removed, we checked the analyses of Frog repeatedly and adjusted the anonymization and pre-processing accordingly (reflected by the dotted arrows in Figure 5.2).

Safe-keeping

After anonymisation, each client was assigned a (random) number. None of the researchers working with the data knew the identity of the client (or the counsellor). The data were stored on encrypted USB-sticks that were only accessible with a password. Incorrectly entering the password multiple times would result in the USB stick deleting all data. To ensure that the e-mail data was not stored anywhere else, the USB-sticks only made contact with a computer when working with the data. The anonymisation and pre-processing were done at Tactus, so no non-anonymised version of the data left their office.

The LIWC program

After pre-processing, we analysed the content of the e-mails making use of the LIWC by Pennebaker, Boyd, et al. (2015), see Figure 5.2 (we give an example in the results section). LIWC consists out of several dictionaries with subcategories (Chung & Pennebaker, 2007; Tausczik & Pennebaker, 2010). For example, *positive emotion words* consisted out of the words *happy*, *pretty*, and *good*; and is a subcategory of *affective processes*. LIWC processed each e-mail, and counted the number of matches in the dictionaries with all words in the e-mails, see Figure 5.2.

If LIWC found a match, the score of that category increased (Pennebaker, Boyd, et al., 2015). For each text file, the output contained 76 variables. All values were divided by the number of detected words in the LIWC dictionary, so the number 0.25 means that a quarter of all the words in the text were also in that specific LIWC category. In addition to these repeated measures, we also calculated one average for each LIWC category (for each client across the four e-mails). We used the Dutch translations of LIWC (Boot et al., 2017; van Wissen & Boot, 2017).

Labelling client data

All the available e-mails were read into Orange (Demšar et al., 2013) making use of the *mail loader* widget (that we developed for specifically for working with e-mail data). With the widgets *mark duplicates* and *remove marked text*, the duplicate texts from e-mails were marked and removed. The widget *corpus viewer* was then used to read the e-mails without the duplicate texts in chronological order. All e-mails were then read so that clients could be labelled as either a '*drop-out*' or a '*completer*', see Figure 5.2.

The clients were labelled as 'completer' when they had received an e-mail with the word *afsluiting* (Dutch for *closure*). These e-mails were inspected to make sure that they were indeed related to a completed treatment which, for example, included finishing the final assignment *actieplan* ('action plan' in *Dutch*). All clients which did not qualify for these criteria, were labelled as 'drop-out'. The labelling of e-mails of was conducted by Van den Hazel (2020).

Drop-out From 770 clients, only 346 completed the full treatment, resulting in a drop-out rate of 55.1%, see Figure 5.1. Because drop-out is our main variable of interest, Table 5.2 is disaggregated to those who completed the intervention ('*completers*'), and to those who dropped-out ('*drop-out*'). The mean number of e-mails written by clients was 20.8, with a maximum number of 116 e-mails.

Data-analysis techniques

We employed three types of statistical models: a logistic regression, a neural network, and decision trees (see Figure 5.2). For all analyses, we reported the precision, recall, accuracy and the F_1-score, and were performed in Python. We analysed the demographic variables and the LIWC counts for the first four e-mails as *repeated measures*, but also as *averages* over these (four) e-mails, see Figure 5.2. The averages consisted out of 76 LIWC variables; the repeated measures out of 76×4 variables. In order to adequately assess model performance while still

maintaining an acceptable training-test-set ratio, we used 5-fold cross-validation. We will only discuss test-set performance (the confusion matrices in the next section only report the test-set numbers).

First, we included all demographic variables in our analyses (gender, age, education, depressive symptoms, reason for starting treatment, baseline alcohol intake, smoking and gambling behaviour). In our second batch of analyses, we also included all text-variables to see what could be gained from the text-analyses. So, we used *two* different versions of the LIWC scores (as repeated measures and as averages), we also conducted *two* types of analyses (models containing only the demographic characteristics, and that in addition also include the text variables), and we employed *three* types of analyses (a logistic regression, a neural network, and a decision tree).

Logistic regression Because the outcome variable consisted out of 'drop-out' and 'completers', a random (or 'naive') distribution of clients would result in a correct classification of around 50%. Because both groups (roughly) had an equal number of observations (see Figure 5.1), we used 50% as our base rate: about half of the sample can be classified correctly with a naive classification. To also have an impression of the 'baseline' performance of statistical models, we first conducted a (standard) logistic regression. For all models, our outcome variable was drop-out, and we used all demographic characteristics and LIWC categories as independent variables. The training method we used for the logistic regression was *mini-batch gradient descent*, with the *binary cross-entropy* as the loss function. See Table 5.5 for an overview of the specific parameters that we used.

Neural network Neural networks are well-known for their predictive accuracy on complex tasks (C. M. Bishop, 2006; Smink, Sools, Tjong Kim Sang, et al., 2020),[2] and are often applied in text mining (Gibson et al., 2016; Hoogendoorn et al., 2017; Ji et al., 2018; Tanana et al., 2015; Xiao et al., 2016). We used a multilayer perceptron, with five fully-connected layers. We used the *rectified linear unit* as the activation function in the hidden layers. The training method was *mini-batch gradient descent*, with the *binary cross-entropy* as a loss function, see Table 5.5 for the specific parameters.

Repeated measures neural network This repeated measures neural network used a slightly different architecture: the first layer of the network only

[2]Smink, Sools, Tjong Kim Sang, et al. (2020) is chapter 4 of this thesis.

Table 5.5: Hyperparameters after fine-tuning.

Model	Eta	Epochs	Batch	Drop	Momentum	Tree depth
Average						
Logistic Regression	.01	100,000	16	.1	0	-
MLP	.05	10,000	16	0	0	-
Decision Tree	.001	10,000	-	-	-	2
Repeated Measures						
Logistic Regression	.01	200,000	16	0	0	-
MLP	.01	5000	64	.5	.2	-
Decision Tree	.005	1000	-	-	-	2
Advanced NN	.01	5000	16	.5	.2	-
MERF Time	-	30	-	-	-	-
MERF Client	-	30	-	-	-	-

takes the demographic data as input, and the LIWC scores only enter the network in the second layer. Maity and Pal (2013) showed that this architecture could improve results when dealing with repeated measures data. The training method was *mini-batch gradient descent*, with *binary cross-entropy* as a loss function, see Table 5.5 for the specific parameters.

Decision tree Decision trees are well-known for the insightful 'decision-maps', that are relatively straightforward to interpret (Breiman, 2001a, 2001b; Kotsiantis, 2013). Ensemble methods are a class of decision tree models that have better predictive performance (Ji et al., 2018), and these so-called boosting methods combine several 'weak' classifiers to improve prediction of the final boosted classifier (Bonab & Can, 2016). We applied XGBoost (Chen & Wojcik, 2016), a (boosted) decision tree that –for many tasks– is known to outperform standard tree-based models (Hoogendoorn et al., 2017). The training method was *batch gradient descent*, with the *hinge* as the loss function, see Table 5.5 for the specific parameters.

Repeated measures random forest We also used two Mixed Effect Random Forests (MERF; 'longitudinal decision trees'). A MERF enhances the standard decision tree by including mixed (or 'random') effects, which can lead to substantial performance improvements when dealing with clustered data (Hajjem et al., 2014). To our understanding, the MERF software of Hajjem et al. (2014) does not include the option to asses both clustered structures simultaneously. Therefore, our first MERF used the repeatedly observed LIWC scores as clusters, and

Table 5.6: Top ten most commonly used words in the e-mails for those who completed the intervention, and for those who dropped out.

Completers		Drop-out	
English	*Dutch*	English	*Dutch*
that	*die*	you (*formal*)	*u*[1]
was	*was*	have	*heb*
he	*hij*	your (*formal*)	*uw*[1]
felt	*voelde*	Dear (*formal*)	*Beste*
glass	*glas*	detox	*detox*
also	*ook*	general practitioner	*huisarts*[2]
counsellor	*therapeut*[2]	are	*zijn*
cancer	*kanker*	kind	*vriendelijke*
it	*het*	use	*gebruik*
pain	*pijn*	your (*plural*)	*jullie*
forum	*forum*[3]	regards	*groet*

Note. This Table presents the results of the unigram model; we discuss the observations numbered 1 to 3 in the *results* section.

as a result, the MERF-model took the longitudinal structure of the data into account. The second MERF used the clients as clusters. The training method we used was *Expectation-maximization*, with the *Generalised Log-Likelihood* as a loss function, see Table 5.5 for the specific parameters.

Results

We first studied the most frequently used words in the e-mails of the drop-out and the completers, see Table 5.6 for an overview of (the translations of) these words. Table 5.7 contains the performance metrics of the models that we employed: we report the accuracy, F_1-score, precision and recall. We assign the most weight to the first two scores, as they balance the other two (and some other aspects of correct and incorrect classifications). Table 5.7 also contains a 'naive' baseline model: it does not base the classification of clients on anything else other than random chance, or a naive assignment of all to the same category.

As can be seen in Table 5.7, the unigram model does not outperform baseline classification. This means that none of the words in Table 5.6 is able to discriminate between the drop-out and completers. In other words, Table 5.6 indicates that it is not possible to identify an e-mail of a drop-out or completer based on the unigrams alone. However, some word classes in Table 5.6 stand out as they could indicate some potentially relevant differences between the drop-out and completers: we discuss these three word classes in the next section.

Table 5.7: The performance metrics of the models.

Model	Accuracy	Precision	Recall	*F1*-score
Baseline				
Negative only	.551	0	0	0
Positive only	.449	.449	1	.620
.449 chance of positive	.505	.449	.449	.449
N-grams				
Unigrams	.591	.538	.633	.582
Average				
Logistic regression	.560	.509	.322	.372
MLP	.575	.568	.272	.346
Decision tree	.610	.562	.638	.594
Demographic only				
Logistic regression	.571	.525	.218	.286
MLP	.570	.525	.293	.356
Decision tree	.587	.540	.570	.551
LIWC only				
Logistic regression	.534	.381	.158	.222
MLP	.529	.458	.208	.283
Decision tree	.579	.523	.717	.604
Repeated measures				
Logistic regression	.595	.533	.721	.609
MLP	.525	.522	.302	.374
Decision tree	.612	.548	.799	.648
Advanced NN	.566	.519	.537	.522
MERF Timing	.541	.533	.474	.501
MERF Client	.525	.514	.471	.491

Word use

Our first observation involves the usage of informal pronouns; in Dutch there is a formal and informal way for addressing others. Three of the most frequently used words by the drop-out are formal (see number 1 in Table 5.6), and were used by the clients to address the counsellor in a somewhat remote manner ("*I don't know what you can do for me, aside from forwarding my file*"). Aside from the 'distance' between the client and counsellor, formal language is also used to express some sort of misconception ("*Did I understand you correctly, you want me to answer all the questions? That will take ridiculously long for you to read*").

Second, the completers often refer to a *psychotherapist*, whereas the drop-out frequently mention a *general practitioner* (see number 2 in Table 5.6). It appears that the completers –in addition to the counsellor from Tactus– also have a psychotherapist ("*Yes, I really believe I need a therapist with whom I can fight about my ideas and thoughts*"). This therapist is often perceived as a source of support ("*My therapist agrees that I could benefit from these situations as well*"), and it appears that the therapist often discusses topics that are similar to the ones in the Tactus intervention ("*I discussed this yesterday as well with my therapist*").

The drop-out on the other hand often mention their general practitioner: the Tactus-counsellors refer the excessive drinkers to a general practitioner ("*I went to my general practitioner, and my blood pressure was good*"). The general practitioner of some appears to be aware of the alcohol dependency ("*My general practitioner knows about my alcohol abuse*"), whereas this is not the case for others ("*I tried to discuss this with my general practitioner, but I was shocked by the reaction I got*").

Thirdly, the completers mention a *forum* (see number 3 in Table 5.6). In addition to the AdB-program, Tactus also offers access to an online internet forum where it is possible to discuss and meet with other participants of the program. According to Postel (2011), the forum receives great user satisfaction, and offers support, motivation, and engagement (p. 136). It appears that the drop-out do not use this forum, as they do not mention it.

LIWC analyses

We analysed the e-mails from the AdB intervention with LIWC (for an overview of the e-mails, see Table 5.4), which resulted in output similar to what we present in Table 5.8. We used these LIWC results as input features for our analyses (see Figure 5.2). We first address the results of the LIWC-averages ('average' in Table 5.7). The confusion matrices corresponding to results of these average results follow in Table 5.9. As the confusion matrices of the LIWC repeated measure

Table 5.8: An example of LIWC2015 output for an example e-mail.

LIWC dimension	*Score*
I-words (I, me, my)	13.8
Social words	10.3
Positive emotions	1.1
Negative emotions	2.3
Cognitive processes	15.5
Summary	
Analytic	19.1
Clout	21.0
Authenticity	94.2
Emotional tone	11.0

Table 5.9: Confusion matrix for the models using the averaged LIWC scores.

Model		*Observed* +	-
Logistic regression	+	44	34
	-	26	29
MLP	+	55	39
	-	15	24
Decision tree	+	53	43
	-	17	20

Note. A completer is labelled with a '+', drop-out is labelled with a '-'. Only the test-set ($N = 134$) has been included.

analyses were similar to Table 5.9, we did not include these here for the sake of brevity.

Table 5.7 does not indicate that any analyses based on the LIWC-averages outperforms naive classification: the accuracy and F_1-score rarely exceed the random baseline classification. For the LIWC averages, the F_1-scores are low for the logistic regression and the multilayered perceptron (MLP in Table 5.7), mainly due to poor recall. The decision tree performed slightly better, with a higher accuracy and a recall that is substantially higher than for the other two

models.

To see if we could improve classification, we also employed longitudinal models that were specifically tailored to the repeated measures structure. For this purpose, we used a repeated measures logistic regression, a (simple and advanced) repeated measures neural network, and repeated measures decision trees (MERF). These results are also displayed in Table 5.7.

Even though we expected an increase in model performance, Table 5.7 indicates that model performance remains similar to the naive classification. The performance of the longitudinal decision tree and MERF are on par with the neural network of the LIWC averages. Even though we included all LIWC categories and demographic characteristics as input in our analyses, Table 5.7 does not indicate that the longitudinal models are (better) capable of predicting drop-out (so for the sake of brevity we did not discriminate between the analyses that did and did not include the demographic variables, mainly because we did not find differences in model performance). Given that we employed a wide array of models, we conclude that there are no 'large', 'powerful', or 'strong' predictors of drop-out in the first four e-mails.

V

Discussion

Web-based psychotherapy is an established alternative to classic face-to-face therapy, with the large drawback that almost all online interventions are plagued by a high rate of drop-out. In our data, we found a drop-out rate of nearly half, which was high, but also expected based on past studies (Kelders et al., 2012). So, *why did these clients drop-out*? We tried to answer this question by testing whether we could differentiate between the first four e-mails that the completers and drop-out sent. We studied these e-mails through a wide array of models, but could not associate the e-mail texts exclusively with either the completers or drop-out.

As some clients drop-out because they already experienced the benefits from the intervention, we conclude that drop-out is a complex and multidimensional construct (see Figure 5.3). Perhaps (LIWC-)dictionaries that target the nuances indicated by Figure 5.3, or dictionaries that are specifically tailored to AUD, will be more helpful in understanding drop-out. Our list of the most frequently used words could be helpful first step, even though these words cannot be exclusively associated with one of the two categories.

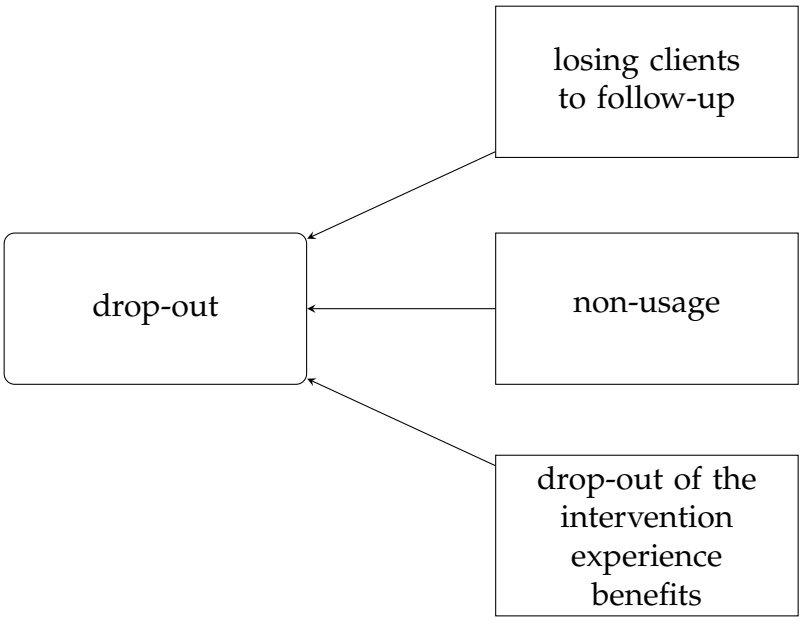


Figure 5.3: Drop-out is not a straightforward proxy of therapeutic success, as it could reflect multiple aspects, and not all are related to therapeutic success.

A qualitative interpretation of the frequently used words

The reason that we are cautiously optimistic is because a qualitative analysis of the frequently used words does indicate that (some of) these words have potential. The (*Dutch*) words '*u*' and '*uw*' –all polite and formal ways to address the counsellor– were used with greater frequency by the drop-out. This could stress the importance of the therapeutic alliance between client and counsellor (G. Clarke et al., 2005; Horvath & Luborsky, 1993). The fact that the completers do not address their counsellors in a formal way could be an indication that they feel they have a stronger therapeutic alliance (Eubanks et al., 2018). The use of formal language could also indicate that it takes some experience to work with a counsellor: it is not straightforward to immediately trust someone with your thoughts. Establishing trust and other aspects of the alliance require effort and become easier with experience (which is perhaps too difficult for the drop-outs to do online). It is possible that the clients who know 'how to work with a counsellor' could be further in their process of becoming less dependent on alcohol, which brings us to the second point.

There is also a difference between the usage of (the word) *therapist* (by the completers) and *general practitioner* (by the drop-out). For some the intervention could be a recommendation from their general practitioner (the unigrams indicate that this group could be the drop-out), while others could have found the intervention on their own because they are further in the process of seeking

help for their alcohol dependency (the unigrams indicate that they could be the completers). This difference in word-use could also indicate that the drop-out could perceive their alcohol usage as a medical problem, whereas the completers could perceive their drinking behaviour to be of psychological nature (and –as a result– they are more open for psychological and psychosocial help). It could be possible that viewing alcohol dependency as a psychological –rather than a medical– problem has benefits (McMurran, 1994).

It is not unlikely that there could be a generational effect as well (but this is speculative): some clients of a certain generation perceive medical care to be the only form of 'real' healthcare, as they do not have much faith in psychological counselling (we have the impression that some older clients could perceive counsellors in general to be '*tree-huggers*', or *geitenwollensokken-types* in *Dutch*). As a result, they could invest less in the therapeutic alliance (as indicated by the use of formal words), because they were less likely to see counselling as a from of help in the first place. Another possibility is that older people simply use more formal language online.

Our last observation is that the Tactus forum could be a great source of support, as only the completers frequently mention the forum. In line with the first observation, it could be that the drop-out are less engaged with the intervention, and try to keep some distance between themselves and the intervention. It could also be that those who actively participate in the forum are further in their own psychotherapeutic process, as they know it also takes some effort from themselves to become less dependent on alcohol. It would therefore not be farfetched to recommend that the intervention could try to establish more tie-ins with the forum (and vice versa).

LIWC is unable to extract information relevant for drop-out from the first four e-mails

Given the popularity of the program, we expected that LIWC would be helpful in determining the textual aspects of therapy that are relevant for drop-out. However, we were unable to achieve satisfactory predictive accuracy based on the features from LIWC. We are confident that we employed many models that suit this data set, but to no avail. We therefore conclude that we were not able to use LIWC to extract information relevant for drop-out from the first four e-mails.

Perhaps LIWC is too 'crude' to pick of the nuances that are present in text data. For example "*I was so angry when I was a child*" is equally counted to the category anger as "*I am so angry right now*", even though the first sentences describes a statement from the past, that might have very well changed by now.

Also, "*I hate my family* " and "*I love my family*" both contain words for the family LIWC category yet have a entirely different emotional connotation. So the category *family* in this example does not allow for a meaningful differentiation between these two statements. As a result, it becomes difficult to target the nuances indicated by Figure 5.3.

The value of e-mail data

Ever since Sigmund Freud introduced the talking cure, conversation is the cornerstone to most forms of psychotherapy. Given the central position of the therapeutic exchange, a careful assessment of the therapeutic language could provide insight into what is happening in therapy. As AdB primarily relies on the exchange of e-mails between client and counsellor, information about the therapeutic processes (drop-out included) should be present in these e-mails. Even though we were not able to produce models that were helpful in discriminating between the drop-out and completers, the qualitative interpretation of our findings does suggest that the e-mails contain relevant information.

Open challenges: strengths and limitations

Web-based interventions offer the opportunity to bring new groups of patients and clients to counselling (Postel, de Haan, ter Huurne, Becker, et al., 2010; Postel et al., 2005), but online drop-out is (often) substantial (Postel et al., 2011). To the best of our knowledge, there are no reports that explore the possibility to study drop-out for AUD specifically through a text-mining approach. Having a better understanding of how (and why) drop-out occurs does not only have scientific value, it can also greatly benefit the clinical practice.

In our study we relied on the LIWC program, but the way we used LIWC did not provide the information that we were interested in. Our choice for a (relatively) 'crude' method such as LIWC does not contribute to a more nuanced and fine-grained understanding of drop-out. Indeed, with respect to LIWC, we present a null-finding. However, we are confident that others will find this to be helpful: as LIWC is a popular tool among psychologists, it is helpful to know the application where it works less well. Also, as we present one of the first attempts to systematically study e-mails for AUD, we tried to go beyond a qualitative analysis of all e-mails, which automatically brought us to LIWC (as it is so well-established). Because there is too much e-mail data available, a complete and 'manual' study of all the aspects from the is too labour-intensive (Snow et al., 2008). The middle ground between rather technical text mining models and a qualitative analysis of the e-mails is the LIWC program.

Also, for a –relatively– small language such as Dutch, LIWC is the best and only readily available alternative. Around 24 million people are estimated to speak the Dutch language. Compared to the English language, that is not a lot. It could be that there are more relevant dictionaries available than LIWC, but that these are only available in English or another language. Simply translating all the dictionaries through standard available translation-software is arguably naive, as it could result in a loss of linguistic and culture nuances (or otherwise context-specific aspects of language).

One of the main limitations is that we worked with a relatively small dataset. However, from the viewpoint of Tactus (that includes counsellor- and client-perspective), it would be reasonable to stress that a dataset such as the one we analysed is quite large, especially in the Netherlands. However, from the perspective of machine learning efficiency, it is also true that the data is *not* large. There are (in total) not as many data-points as one would prefer for machine learning. It is indeed possible to make two equally valid argumentations, that would result in the conclusion that the data is both *small* and *large* at the same time (stressing how context dependent these terms are).

Future research

Our first recommendation for future research is in line with one of the main limitations of our work (we intend to give some general recommendations): include the nuances of drop-out. It took us a lot of time to clean and pre-process the data for the analysis, and it would take even more time to prepare and label the data to pick up this nuance as well. The AdB intervention ran for a couple of years: if it would have been easier to perform some simple text data-analysis during the intervention, we are confident that this nuance would have come to light early on. If counsellors could then mark the specific cases of drop-out, and it would become straightforward to explore drop-out in a more meaningful way. TPCR related studies would therefore benefit from *structured* datasets that include more labelling from the counsellor, so that it becomes possible to conduct a more nuanced analysis of the text constructs. It should be possible to decrease the time spend on pre-processing, while the value of the data analysis increases greatly.

Developing systems that are more suitable for the export of the (relevant) texts from e-mails is important for numerous reasons. As of yet, more often than not this requires some familiarity with programming, which is –to the best of our knowledge– not part of the standard curriculum of psychologists trained at the graduate level. For the clients it was also natural to include quotes from

the counsellors in the data, but that posed some challenges in analysing the data. If it would have been possible to perform some simple text analyses during the intervention, it would have been easy to spot this as problematic. Some sort of labelling during the intervention would have been helpful here as well.

Conclusion

Web-based interventions are effective, but are also plagued by a high rate of drop-out. For alcohol dependency, there is an astoundingly large gap between first treatment contact and the onset of problems. Often, online help is sought earlier, but alcohol dependency is typically associated with high drop-out. Nowadays, there is a plethora of new research methods available, which re-establish the importance of understanding the (underlying) change processes. As traditional forms of psychotherapy are increasingly complemented by online interactions between client and counsellor, there is also data available for these purposes. As the active ingredients of therapy are included in the exchange of several e-mails, this verbal exchange should contain a wealth of information on the psychotherapeutic change process. Although we did not find effects of considerable size, we are confident that our manuscript is of value to future studies that explore drop-out through the linguistic features of e-mail data.

Appendix: Case descriptions

To further characterize the clients that were included in our sample, we give four anonymised case descriptions. All case descriptions rely on the information that respondents reported about themselves on the assessment questionnaire, and include some of the things they wrote about themselves in the e-mails. The case descriptions come from Giesler (2019) and Krstić (2019).

Case 1 A 38 year old female, who regularly consumed alcohol for over ten years. At the onset of the treatment, she consumed 62 units of alcohol per week. Her goal of participating was to reduce alcohol intake to prevent health problems from reoccurring, and to make progress in her studies. The main motivation for drinking includes its rewarding and relaxing effect. Previous drinking habits have caused her loss of memory, missing lectures, self-sabotaging patterns (in terms of diet), headaches, and sleeping problems. An important factor for this client is also her husband, with whom she drinks on many occasions. The client noticed that she had problems quitting even when others around her were not drinking.

Case 2 A 49 year old male, living with his wife and two children. The level of consumption at the beginning of the treatment included between 20 and 50 units per week. The client set the goal of reducing consumption, but emphasized the social and practical factors that justified the drinking behaviour. The client reported several physical problems including memory loss, tiredness, decrease in sexual arousal, and depression. His wife is an important factor during the treatment, and is described as a "*regular drinker*". During the treatment, mood swings were dominant. Alcohol was important for the job of the client, as it was an important contributor to closing (business) deals. Mood swings were also frequently reported in the e-mails, emphasizing that alcohol is perceived as a reward.

Case 3 Client is a 27 year married female, who got pregnant during the treatment. The client had a medical history of potential eating disorders (she reported binge eating, dieting, and excessive exercises). Physical symptoms include fatigue, headaches, and sleep deprivation. At the beginning of the treatment, the client consumed 32 units of alcohol per week. Alcohol is perceived as a rewarding, socializing, and self-empowering factor. Due to her working environment, the client feels that she cannot deny drinks that are offered to her. On an emotional level, the client referred to disappointment and frustration, which

are related to her self-esteem problems that are also reflected in her problems with body-image and previous binge eating episodes. For this client, establishing control over her symptoms are of great importance. Due to pregnancy, the client decided to alter her goal from reduction to complete abstinence.

Case 4 A 48 year old male, living alone and has a conflictual relationship with his sister. His mother is in a nursing home, and he often interacts only with one friend. The initial consumption included 105 units per week. The client reported having a breakdown, due to which he consumed blood pressure and cholesterol medication. The most dominant emotions in the e-mails were anxiety, frustration, and overall dissatisfaction with his relationships, work, and life. The client showed indications of suicidal ideation and is often apologetic towards the counsellor. He perceived alcohol as a social factor, highlighting that "*there is no life or parties without alcohol*."

What
Works
When
for
Whom

6 Understanding Therapeutic Change Process Research through Multilevel Modelling and Text Mining

Abstract

Nowadays, traditional forms of psychotherapy are increasingly complemented by online interactions between client and counsellor. In (some) web-based psychotherapeutic interventions, meetings are exclusively online through asynchronous messages. As the active ingredients of therapy are included in the exchange of several e-mails, this verbal exchange contains a wealth of information about the psychotherapeutic change process. Unfortunately, drop-out related issues are exacerbated online. We employed several machine learning models to find (early) signs of drop-out in the e-mail data from the 'Alcohol de Baas' intervention by Tactus. Our analyses indicate that the e-mail texts contain information about drop-out, but as drop-out is a multidimensional construct, it remains a complex task to obtain accurate predictions. Nevertheless, by taking this approach, we present insight in the possibilities of working with e-mail data and present some preliminary findings (which stress the importance of a good working alliance between client and counsellor, distinguish between formal and informal language, and highlight the importance of Tactus' internet forum).

Keywords: Therapeutic Change Processes Research (TCPR), Multilevel Models (MLMs), text mining, process data, web-based interventions, text variables

Smink, W. A. C., Fox, J. P., Tjong Kim Sang, E., Sools, A. M., Westerhof, G. J., & Veldkamp, B. P. (2019). Understanding Therapeutic Change Process Research through Multilevel Modelling and Text Mining. *Frontiers in Psychology*, *10*, 1186. https://doi.org/10.3389/fpsyg.2019.01186

Introduction

Traditional forms of psychotherapy are nowadays increasingly supplemented by online interactions: it is not uncommon that a counsellor seeks contact with a client through e-mail, text, chat, or other text-bearing messages. As the contact between counsellor and client becomes increasingly digitally mediated, it should be possible to trace the factors that contributed to the beneficial outcome of treatment back to these textual interactions.

In this light, the field of *Therapeutic Change Process Research* (TCPR) re-establishes its importance. TCPR aims to identify the mechanisms through which psychological treatments bring about positive and therapeutic change (Elliott, 2010, 2012; Greenberg, 1986; Orlinsky et al., 2004). TCPR has a long-standing tradition of studying the linguistic 'products' of therapy (e.g., homework exercises, diaries, transcripts) in order to understand therapeutic change (Imel et al., 2015; Kazdin & Nock, 2003).

The rising popularity of Internet-based interventions (cf. Hoogendoorn et al., 2017) allow researchers to ask new TCPR research questions and re-establish the relevance of several known questions. Questions pertaining to the change processes that are beneficial to clients necessitate investigation of the 'active ingredients' of therapy, of which many are linguistic (Imel et al., 2015; Muntigl & Horvath, 2005). TCPR has thus the potential to reveal the fundamental processes that are related to change. Aside from insight in what helps patients improve their functioning and reduce (clinical) symptoms, the importance of TCPR is also related to the rising number of people diagnosed with mental health disorders (Andrade et al., 2013; Whiteford et al., 2013).

Over many decades, researchers attempted to answer TCPR questions; Orlinsky et al. (2004) estimated that there are more than 2000 published process-outcome studies of psychotherapy. Crits-Christoph et al. (2013) discuss several (methodological) issues related to TCPR, and express that "*individual psychotherapy is not based just on an individual: it is a dyadic relationship consisting of a patient and therapist*". Similar to Kenny and Hoyt (2009), Crits-Christoph et al. (2013) argued that –from a statistical point of view– patients are nested within their therapist, hence, TCPR is concerned with *multi-level models* (MLMs; also known in the literature as hierarchical linear models, mixed models, random coefficient, or random effects models).

Yet, we found few studies that applied MLMs specifically to study therapeutic language. In the current work, we will present an approach for the study of therapeutic change processes based on text mining and MLMs by (re-)analysing e-mails send between counsellor and client (Lamers et al., 2015). We do so by

first making a comprehensive argument for the importance of understanding multi-layered change processes (Knobloch-Fedders et al., 2015), and argue for the use of text mining to study TCPR.

TCPR: Therapeutic Change Process Research

Progress in psychotherapy research is not made by only demonstrating the (average) effectiveness of a treatment; the history of psychotherapy research is marked by a gradual increase in the understanding of psychotherapeutic change processes (Braakmann, 2015; Orlinsky et al., 2004). Hence, psychotherapy benefits from a greater understanding of TCPR,[1] which is defined as the scientific investigation of what occurs during psychotherapy, with regard to its clinical meaningfulness; in other words, it investigates the process through which clinically relevant changes occur within psychotherapy (Gelo & Manzo, 2015, p. 248).

Questions concerning the underlying processes that benefit the client also align with the interests of many clinical practitioners (Norcross & Wampold, 2011): *what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of *circumstances* (Paul, 1967; Tasca et al., 2015, p. 111)? Studies aimed at demonstrating average effects at group level fail to show what aspects of the intervention are related to the change the intervention realized (Barkham et al., 1993; Nock, 2007). Still, more effort is devoted to the analysis of the outcomes of psychotherapeutic interventions.

TCPR and the study of the therapeutic conversation

As early as Freud's talking cure, the importance of looking at language to understand the therapeutic process has been recognized. Conversation is still the interactive medium central to most forms of psychotherapy (Muntigl & Horvath, 2005). The idea that the verbal exchange between counsellor and client contains important ingredients of therapy fuelled TCPR (Elliott, 2010; Greenberg, 1986; Hill & Lambert, 2004), which is known for its a long-standing tradition of studying the linguistic 'products' of therapy (e.g., homework exercises, diaries, transcripts) in order to understand therapeutic change (Gelo et al., 2015, p. 303, 392).

[1] It should be noted that various terminologies are used in the literature, e.g., Change Process Research (CPR: Elliott, 2010; Greenberg, 2007), Psychotherapy Process Research (PPR: Gelo et al., 2012), and some of the early works simply refer to 'change' (Braakmann, 2015; Hill & Corbett, 1993). The term 'process-outcome research' is also often used, for example by Orlinsky et al., 2004, who defined it as "(primarily) the actions, experiences, and relatedness of patient and therapist in therapy session when they are physically together, and (secondarily) the actions and experiences of participants specifically referring to one another that occur outside of therapy sessions when they are not physically together" (Crits-Christoph et al., 2013, p. 311). To emphasize that we are dealing with change resulting from *therapy*, we propose to describe change processes as TCPR: *Therapeutic Change Process Research*.

For example, the *Narrative Processes Coding System "focused on the strategies and processes by which a client and counsellor transform the events of everyday life into a meaningful story that both organizes and represents the client's sense of self and others in the world"* (Angus et al., 1999).

Another reason to specifically choose text over other types of TCPR data is that a valid understanding of psychotherapeutic processes require measurements collected from multiple perspectives, including that of the client, counsellor, and (possibly) external observers (Knobloch-Fedders et al., 2015). A good way to do so is to study the text-based representation of the therapeutic interaction. Because these transcripts are a direct observation of the therapeutic process, they reflect what actually happened in therapy. Transcripts can thus provide the basis to obtain the perspective of the client, counsellor and (or independent) observer on the therapeutic process. Such interpretations are usually measured by questionnaires or interviews, and are retrospective reflections. Transcripts come with the additional benefit that they are relatively straightforward to obtain after providing transcripts of the therapeutic conversation.

In this light, it is not unsurprising to see TCPR moving towards web-based interventions (a variety of different terms is used for –essentially– the same concept; Oh et al., 2005). Aside from being cost-efficient, web-based self-help interventions directly produce the textual interaction between therapist and counsellor, and come with the additional benefits that they are effective (Andersson et al., 2014; Andrews et al., 2010; Andrews et al., 2004), and easily accessible by large groups of people (Barak et al., 2008; Hoogendoorn et al., 2017; Wang et al., 2007). Just like transcripts, assessment of the interaction between counsellor and client in a web-based intervention has the potential of being a direct observation of the therapy process (Elliott, 2012; Gelo et al., 2012; Pennebaker et al., 2003; Schegloff, 2007).

Transcription and manual analyses mark the labour-intensive nature of TCPR, which is also the main reason why the field did not yet reach its full potential (Smink, Sools, van der Zwaan, et al., 2019).[2] Traditional research methods start with the recording and transcription of a psychotherapeutic intervention so that human raters can (manually) code and analyse these transcripts (Atkins et al., 2014). Because the understanding of change processes mainly relies on qualitative analysis, these methods are only as fast as the researcher(s) conducting the research, which in practice limits their use to small scale studies (Atkins et al., 2012; Imel et al., 2015).

To strike a balance between TCPR's ambition to unravel the black-box through

[2]Smink, Sools, van der Zwaan, et al. (2019) is chapter 2 of this thesis.

which therapy attains its effects and the labour-intensity of the TCPR methods, we propose to use automated text analysis methods. Text mining, a computational approach to text analysis, can be used to automatically extracted text features that can contribute to the understanding of the active ingredients of therapy. We are observant of the criticisms that algorithms have yet to achieve the same exploratory depth of analysis as humans. However, in our view, it would be a shortcoming to TCPR's ambitions if the insights that basic text features can offer remain unused. In the next session we will discuss how text mining can scale up TCPR by finding text-based predictors –also known as input variables or independent variables– from therapy related texts. We will do so making use of *multi-level models* (MLMs), an advanced statistical model that is able to capitalize on the hierarchical structure of text data.

Text mining: scaling up TCPR

As language is an important mediator of psychotherapeutic processes, obtaining information about these processes through texts is one of the first applications of text mining. Mergenthaler (1996) compared five computer-assisted measures for the analysis of textual data of two psychotherapies, and was among the first to apply text mining for psychology. He used text mining, which he then called "computer assisted analysis of textual data", to identify turning points in sessions, which could then be explored more deeply by humans through (qualitative) analyses methods. Anderson et al. (1999) developed *Computer Assisted Language Analysis System* (CALAS) to examine the relationship of various linguistic measures to outcome measures in high and low verbalized affect segments. Many applications of text mining are still centred around finding key moments in the therapeutic process (cf. Fontao & Mergenthaler, 2008; Lepper & Mergenthaler, 2005; Pfäfflin et al., 2005), which is also a common approach in TCPR (e.g. the '*Significant Events Approach*' in Elliott, 2010).

Practically, typical text mining approaches in psychology include counting words, identifying topics, and coupling the terms to a domain-specific ontology (Hoogendoorn et al., 2017). Text mining[3] refers to a general methodological framework that includes several automated methods to analyse large corpora of texts (cf. Jurafsky & Martin, 2017). As text mining is a methodological framework that combines and includes numerous techniques and methods from many disciplines, it is not surprising that terms referring to the automatic extraction of

[3]We recommend R. Feldman and Sanger (2007) and Jurafsky and Martin (2017), and Manning and Schütze (1999) for a detailed overview of text mining. For aspiring text mining practitioners, we recommend the NLTK library available in the programming language Python (Bird et al., 2009).

information from text are used sometimes interchangeably, such as text mining and NLP.

Text mining emotions

The *Linguistic Inquiry and Word Count* (LIWC) software by Pennebaker, Boyd, et al. (2015) is used by many researchers, and has showed to be effective in predicting therapist empathy (Gibson et al., 2015), counsellor behaviour (Pérez-Rosas et al., 2017), and identifying emotional and cognitive process in psychotherapy (McCarthy et al., 2017). LIWC categorizes word usage by counting the percentage of words that reflect –among other categories– thinking styles, emotional states, and social concerns (Hirsh & Peterson, 2009; Pennebaker, 1997; Tausczik & Pennebaker, 2010). LIWC taps into the underlying idea that word use is one of the most direct means of expressing thoughts and feelings (Fast & Funder, 2008), as the way individuals talk and write provides a window into their emotional and cognitive worlds psychological characteristics (Pennebaker, Francis, et al., 2015; Pennebaker et al., 2003).

The writing intervention by Lamers et al. (2015) focused on different life themes, with one theme central to each of the seven modules. By asking clients to describe specific positive and several difficult memories, clients adjusted their life stories step-by-step by integration of these memories. Lamers et al. (2015) did not include the content of the e-mails in their study.

Previous studies showed that positive therapeutic outcomes from writing interventions are associated relatively high rate of positive emotion words, few negative emotion words, and with an increasing number of 'cognitive' words[4] throughout the intervention (e.g., R. Campbell & Pennebaker, 2003; Pennebaker, 1997; Pennebaker, Francis, et al., 2015). As the intervention of Lamers et al. (2015) focuses specifically on positive and difficult memories and emotions with the aim of integrating these two, we study words the reflective of these aspects in e-mails. As the intervention by Lamers et al. (2015) aims to improve integration of positive and negative memories, we expect that LIWC's 'cause' and 'insight' categories are mostly reflective of that process. We aim to find further evidence for these findings in data from Lamers et al. (2015), by relying on text mining and multi-level models.

Text mining and MLMs

Although the idea to relate words or textual aspects (in psychotherapeutic texts) to outcomes is well-established in TCPR, there are methodological issues that

[4] *Cognitive words* are a word category from the LIWC program.

are specifically relevant when analysing text data. Studying change processes in e-mails mandates accounting for the *dyadic* relation (Crits-Christoph et al., 2013, p. 301), and is therefore dependent on both the counsellor and client.

While the assumption of *independence of observations* is the basis for traditional statistical models, such as the ANOVA or regression model, some text mining models relax this assumption. For example, the *naive Bayes* classifier *assumes* independence assumptions between observations. The model classifies units to the category that has the highest probability; a common application of the model is the spam-filter, where e-mails are classified as either spam or 'ham' (no-spam). He et al. (2012) used *naive Bayes* to find words that could discriminate between texts written by soldiers with or without PTSD.

Naive Bayes is a family of algorithms based on the assumption that the value of a particular (text)feature is independent of the value of any other feature. This independence assumption is too strong ('naive'); in reality, independence does not hold for texts that are written by the same person. In doing so, the model 'naively' neglects the nesting of e-mails within person, ignoring the assumption of independence. In the next section work, we will argue for the importance of applying MLMs to analyse textual data for correct statistical inference, as MLMs do not violate the non-independence in e-mail data (Kenny et al., 2002). A consequence of failing to recognize the nested and hierarchical structures in e-mails is that standard errors of the estimated coefficients are underestimated, leading to an overstatement of statistical significance. MLMs recognize the existence of hierarchies in data by allowing for residual components at each level of the hierarchy.

Psychotherapy as a multi-levelled procedure

Because MLMs offer the possibility to include predictors at the level of the individual, the group and at any other level of organization, the model arises quite naturally for TCPR (Raudenbush & Bryk, 2002a). Many individual change phenomena can be represented through a two-level hierarchical model. The first level represents each clients' development by an individual growth trajectory that depends on the repeated measures for each client. The second level unit represents variables that are not repeatedly measured, such as gender, income, or depressive symptoms. The first level consists out of –for example– experienced pain at the beginning, middle, and at the end of therapy. The second level consists of the clients themselves, who could be (at a third level) nested within their therapist, for examples see Baldwin and Imel (2013), and Baldwin et al. (2007).

From a statistical viewpoint, TCPR practically equates to research questions concerning either a (longitudinal) development over time (Adler, 2012; Baldwin et al., 2007; Crowder & Hand, 1990; Fitzmaurice et al., 2011; Nissen-Lie et al., 2010), an (dyadic) interaction between a counsellor and its client (Crits-Christoph et al., 2013; Kenny & Hoyt, 2009; Tasca & Gallop, 2009), or to both. MLMs are –compared to traditional statistical methods– particularly useful to both of these situations as they capitalize on hierarchically organized data. Many kinds of data, including observational data collected in the human and biological sciences, have a hierarchical or clustered structure.

Considering that the psychotherapeutic practice is a multi-levelled procedure, it becomes apparent that client and counsellor are the two pre-eminent levels of organization. As counsellors (almost) always treat more clients, clients could be viewed as grouped within their counsellor, similar to the students being *nested* within their class (Crits-Christoph et al., 2013; Kenny & Hoyt, 2009). Crucial to any MLM is that the unit of analysis at the lowest level (the students or clients) are nested within higher level units (classes or counsellor), that itself could also be nested within (higher) even higher units (schools, therapeutic practices, or clinical institutions).

Many of the applications of MLMs in psychotherapy resolve around the question of how to assess psychotherapeutic effectiveness. Adelson and Owen (2012) examined the influence of psychotherapists on clients' clinical outcomes. Baldwin et al. (2007) and Marcus et al. (2009) both showed that higher rates of therapeutic alliance could be relate to better therapeutic outcomes through MLMs (Crits-Christoph et al., 2013). Baldwin and Imel (2013) searched the literature for studies comparing outcomes of therapists. Nissen-Lie et al. (2010) accounted for variation in early patient-rated alliance by means of various self-reports of therapists providing treatment in a naturalistic outpatient setting.

Research questions

Online a client is treated essentially through the language their counsellor uses, therefore the verbal interaction contains many important ingredients that bring about change. TCPR faces two challenges: first, how to derive meaningful change processes from (the) large bodies of texts (that web-based interventions produce)? Second, how to assess these complex, varied and multi-layered processes? These two questions are intimately linked: insight in complex change processes gives an indication of how to derive other meaningful processes, and visa-versa.

We therefore advocate the combination of text mining and MLMs: the former

offers tools and methods to discover patterns and trends in texts; the latter can analyse processes that vary at multiple levels. As the study by Lamers et al. (2015) is a writing intervention of which the writing assignment, the e-mails themselves, and the outcomes of the intervention are available, we give a proof-of-concept based on data from this study.

Method

Participants

The dataset derived from 174 clients who were recruited by Lamers et al. (2015) through advertisements in Dutch newspapers and websites. Only participants who felt depressed and were interested in writing about their life were included by Lamers et al. (2015). The sample was thus a self-selected group of individuals who had expressed interest in the program.

All participants had moderate depressive symptomatology and were randomly allocated to either the life-review '*the stories we live by*' (auto-biographic writing; AW), or the '*expressive writing*' (EW) intervention, or a waiting list condition. The mean age of the participants in the AW condition was 57.7 ($SD = 10.3$) years old, and the majority was female (75.9%). The mean age in the EW condition was 56.8 ($SD = 7.9$), and the majority was female (77.6%). In both conditions, the majority of the participants received a higher form of education (i.e. universities or colleges; AW: 48.3%, EW: 37.9%). For more details see Lamers et al. (2015).

Design: Study by Lamers et al. (2015)

We discuss the design of the current study in two sections. First, we discuss the design of the study by Lamers et al. (2015), then we discuss the design of the current study.

Auto-biographic Writing (AW) The AW condition was a life-review self-help intervention that consisted of homework assignments, divided over modules that had to be completed over the course of ten weeks. Clients communicated about their progress with trained counsellors through a weekly e-mail interaction. According to Lamers et al. (2015) the self-help model program was based on insights from the autobiographical memory (Brewin, 2006; Serrano et al., 2004; Williams et al., 2007), narrative therapy (White, 2007; White & Epston, 1990), and life-review (Birren & Deutchman, 1991; Bluck & Levine, 1998; Butler,

1963; Haight & Webster, 1995; Westerhof, Bohlmeijer, & Webster, 2010), and has been shown effective in previous studies (Korte, 2012; Westerhof et al., 2017).

Expressive Writing (EW) According to Lamers et al. (2015) the EW intervention was based on the method of expressive writing (Pennebaker, 1997). The method consisted of daily writing about emotional experiences, for $15 - 30$ minutes on $3 - 4$ consecutive days during one week. Lamers et al. (2015) extended and adapted this method to an intervention with seven modules, to make it a comparable with the life-review intervention.

Design: Current study

Our first intention was to demonstrate how text mining can be used to obtain change processes from e-mails. Lamers et al. (2015) concentrated their efforts on the analysis of the outcomes of the interventions but did not analyse the content of textual characteristics of the e-mails. After pre-processing, we obtained the insight, cause, positive and negative emotion words from the LIWC program.

Our second intention was to demonstrate how multi-level models (MLMs) can be used to assess text-based measures of e-mails to aid understanding of the change processes. Similar to Lamers et al. (2015) we used the post-treatment measurement of the CES-D scale as the main outcome variable.

Materials

Questionnaires

The data available to us included the pre- and post-therapeutic measurements of the CES-D. The *Center for Epidemiologic Studies Depression Scale* (CES-D) is a brief self-report questionnaire to measure severity of depressive symptoms in the general population (Radloff, 1977). Lamers et al. (2015) used the Dutch version of the CES-D (Beekman et al., 1997); higher CES-D scores indicated more depressive symptoms (20 items, range $0 - 60$, $\alpha = 0.78$).

The intervention of Bohlmeijer and Westerhof (2010) teaches participants about autobiographical reasoning by specifically improving the ability to reason about the autobiographical self (Lamers et al., 2015). This form of reasoning describes the process of relating episodic memories to the conceptual self (Pasupathi & Carstensen, 2003; Thorne et al., 2004). By making the moral of an individual's life-story explicit, (s)he obtains insight in what the particular memory could reveal, explain, cause, give insight, or provide a (life) lesson learned about the (autobiographical) self. These processes are extensively researched by

–for example– Pennebaker and Chung (2011), mainly in the context of showing how analogue experiences, such as emotions, are translate to digital forms that bear meaning, such as of stories.

This process is operationalized by phrases that LIWC analyes can detect from the *insight* (e.g. "I now realize that . . . ") and *cause* (e.g. "I understand why . . . "; Pennebaker & Chung, 2011). As the increase in insight and cause words are intricately related to emotional writing, we also study the (increase in) positive words, and (decrease in) negative words from LIWC (Pennebaker & Chung, 2011; Westerhof, Bohlmeijer, van Beljouw, et al., 2010).

Software

We used the LIWC software of Pennebaker, Francis, et al. (2015) to analyse the e-mails for the emotion and insight categories. We used the `NLTK` library of Bird et al. (2009) in the programming language `Python` (Python Software Foundation, 2020, version 3.6), for pre-processing and dividing the e-mail texts in words and sentences.

For our statistical analyses, we relied on the programming language `R` (R Core Team, 2020, version 3.5.1). We used the `lme4`-package for estimation and evaluation of our MLMs (D. Bates et al., 2015), and the `psych`-package for making descriptions of our variables (Revelle, 2018).

Data

The data included the pre- and post-therapeutic measurements of the CES-D scale, and the e-mails exchanged between counsellors and clients (2079 e-mails in total).

Complete cases

In total, data of 174 clients was available to us from Lamers et al. (2015). We only used clients with no missing data. 166 of the 174 clients (95.4%) had a complete CES-D score. Not all e-mails were available, so we could only analyse the e-mails of 104 clients (59.8%). After removing duplicates, we included 97 clients in our analyses (55.7%, all percentages calculated against the original total of 174 clients).

Anonymization

Identifying information has been removed from the dataset that contained the outcomes ('structured' data), we identified clients based on a unique four digits

number. The e-mails ('unstructured' data) have been anonymized by removing all (e-mail) addresses, phone numbers, names of persons, organizations, and locations. Client names and counsellor names have been replaced by the previously mentioned unique four digits number so that it remained possible to identify which mails were written by the same person and which clients were treated by the same counsellor. The counsellors were also anonymized.

Process data

The e-mails of Lamers et al. (2015) should include the whole therapeutic process because they are the only form of interaction between counsellor and client. The e-mail procedure is explained (in Dutch) in detail in Bohlmeijer and Westerhof (2010). We will give some quotes that we translated from Dutch to English to give an impression of process data in a therapeutic context.

The first quote comes from a female participant: "*My trust in people is damaged pretty badly, I'm no longer in such good faith as I was in the past*". In response, the counsellor asks: "*Can you tell us a bit more about this? How did this happen? Are there times when you feel that you can trust people?*".

In the second week a male participant writes: "*By writing about myself, and especially naming the nice aspects about my life, I notice that writing is already paying off*". In the sixth week he writes: "*I feel that I am coming back to who I am*". He also expresses his graduate towards the counsellor: "*I do not have a specific question for you, a reaction from you based on my writing already is already enough. However, if you do ask questions, that would help me even further*".

The third example comes from a (different) female participant: "*How should I continue with my life? Is it okay? Almost thirty years ago I lost my brother and my sister-in-law. I lost my 10-year-old daughter…Losing a child is pretty much the worst thing that can happen to you*". In week seven she wrote (about her daughter): "*The tears are rolling down my cheeks as I think about you intensively. Over the duration of the course I have learned to balance between positive and negative emotions by means of communication or through writing. I succeeded, because I know that you knew that I am still an optimist in life. You and dad have a share in this. You were both never judgemental, but always stimulating*".

Procedure

Selection of the text variables from LIWC

We chose to use the number of *insight* and *cause* words from the cognitive process category, and the number of *positive* and *negative* words form the LIWC program

(Pennebaker, Francis, et al., 2015). We had several reasons for doing so, first of all, past studies showed that positive therapeutic outcomes are associated with writing assignments of individuals that include relatively high rates of positive emotion words, few negative emotion words, and with an increasing number of cognitive words throughout the intervention (L. F. Campbell et al., 2013; R. Campbell & Pennebaker, 2003; Pennebaker, 1997; Pennebaker, Francis, et al., 2015). Secondly, these basic text features are –as the name implies– relatively straightforward to obtain from an e-mail. Third, it is our ambition to show how textual information can be *obtained* through text mining and *analysed* with MLMs. We do not aim to advance TCPR theory in our current paper: determining which textual predictors are meaningful is beyond the scope of our work. We intend to show how TCPR can be modelled in e-mails. Lastly, by bridging text mining and MLMs other TCPR researchers are enabled to to advance TCPR theory using these two methodologies.

Pre-processing

We used the NLTK library to preprocess the e-mails. NLTK counts sentences by counting word-terminal end-of-sentence punctuation like the period, question mark and / or exclamation mark. NLTK has a limited list of abbreviations, which are not included in the punctuation/sentence count. Word-internal punctuation, like the first period in "e.g.," is ignored. Handling of interjections depends on their punctuation, for example, "Oh?" is a separate sentence while "Oh," is part of the following sentence. Sentence fragments and quotes with end-of-sentence punctuation are counted as separate sentences.

NLTK is an often used Python library for text pre-processing, as it provides detailed documentation in Bird et al. (2009) on the order and content of the preprocessing steps.

Pre- and post-therapeutic measurements of the text-variables

We calculated the pre- and post-therapeutic scores of the text-variables (*insight*, *cause*, *positive* and *negative* words form LIWC program) by averaging over the number of these words as counted by Pennebaker, Francis, et al. (2015) in the first and last three e-mails of the intervention by (Lamers et al., 2015). The original intervention also included a third time-point (T_0 a depression measure at the onset of the writing treatment, T_1 a measure at the end of the treatment, and T_2 a follow-up measure). However, only for the first two measurements (those at the beginning and end of therapy) we had e-mail data available. Hence,

we dropped the follow-up measure (T_2) from our dataset, as we could not use in our text mining models.

Analyses

In total, we estimated five MLMs, see Figure 6.1 for an overview and the `R` code. The regression equations below will give an indication of how the `R` code and equations are related. The data we used were the pre- and post-therapeutic measurements of the CES-D and the *insight*, *cause*, and the *positive* and *negative emotion* words of the LIWC (Pennebaker, Francis, et al., 2015).

The pre- and post-therapeutic measurements of the CES-D scale were considered to be an outcome variable of the MLMs. Each MLM had a random intercept for the client to describe the variability in outcome scores across clients. An index i is used to refer to a pre-therapeutic score ($i = 1$) or post-therapeutic score ($i = 2$), and an index j is used to refer to the j^{th} client. Then, the outcomes can be described with a MLM, which is represented by

$$\text{CES-D}_{ij} = \mu + u_{0j} + \mathbf{X}_{ij}\boldsymbol{\beta}_1 + e_{ij}. \tag{6.1}$$

The errors e_{ij} are assumed to be normally distributed with a mean of zero and variance σ_e^2, and the random intercepts u_{0j} is also assumed to be normally distributed with mean zero and variance τ. The parameter μ is the general mean across scores. The predictor variables are stored in a matrix $\mathbf{X}$.The common effects, $\boldsymbol{\beta}_1$, represent the effects of the predictor variables on the outcomes CES-D. The predictor variables $\mathbf{X}$ explain variance in scores across the pre- and post-therapeutic measurement, and do not explain any change between the pre- and the post-therapeutic scores.

To assess change, an indicator variable is used for the post-therapeutic measurement with $D_{1j} = 0$ for all the pre-measurements, and $D_{2j} = 1$ for the post-measurements. A significant interaction between the post-therapeutic measurement scores and a predictor variable would identify a change. The MLM described in Equation (6.1) can be recognized as a repeated measures model, where the model describes the profile of two measurements for each subject. The well-known models for pre- post-therapeutic measurements are the change-score model (the difference in outcomes is regressed on the predictor variables) and the regressor variable method (ANCOVA; the post-therapeutic measurement is regressed on the pre-therapeutic measurement and predictor variables, e.g. Allison, 1990).

Allison (1990) and Kutner and Brogan (1982) showed that the repeated measures model is more general than the change score model, which is more restric-

tive and provides less information about the data. Furthermore, it is possible to control for additional group differences at the pre-therapeutic measurement by including additional predictor variables (Schmidt et al., 2016). This can be beneficial for instances when different groups have not been randomly assigned to different treatments and pre-therapeutic measurement differences between groups need to be accounted for to measure treatment effects. According to the repeated measures model, the MLM for the CES-D scores using the post-therapeutic measurement indicator D is given by,

$$\begin{aligned} \text{CES-D}_{1j} &= \mu_0 + u_{0j} + \mathbf{X}_j\boldsymbol{\beta}_1 + e_{1j} \\ \text{CES-D}_{2j} &= \mu_0 + u_{0j} + \mathbf{X}_j\boldsymbol{\beta}_1 + \mu_1 I(D_{2j} = 1) + \mathbf{X}_j I(D_{2j} = 1)\boldsymbol{\beta}_2 + e_{2j}. \end{aligned}$$

The parameters $\boldsymbol{\beta}_1$ represent the common effects of the predictor variables $\mathbf{X}$ on the outcomes CES-D and explain part of the common variance in the pre- and post-therapeutic measurements. The intercept μ_0 represents the average score level at the pre-therapeutic measurement, and the μ_1 the average change in scores between the pre- and post-therapeutic measurements. Given the effects of the predictor variables, the μ_1 represents the assessed average change in measurements that is not explained by any predictor variable. The parameters $\boldsymbol{\beta}_2$ represents the contribution of the predictor variables in explaining unique variance in the post-therapeutic measurement scores. Significant interaction $\boldsymbol{\beta}_2$ effects identify and explain a change in scoring between the pre- and post-therapeutic measurements.

The first model, our 'null' model, acted as a baseline, hence the name Mo. In Mo, we test whether a random intercept for each client explains variability in outcome scores across clients. In ME1, we test whether the text-predictor variable positive emotion words contributes to explaining the unique variance in post-therapeutic scores. In ME2, MCP1, and MCP1 we test similar hypotheses, but then with the number of the number of negative emotion, insight and cause words.

Results

We intended this section as a guideline for TCPR researchers who aspire to use text mining for multilevel modelling. We start with a statistical summarization of the variables that we used in our five multilevel models. Then we present and interpret the fixed and random effects of these models, and the corresponding goodness of fit measures. In doing so, we hope to give guidance of how these

Table 6.1: Descriptive statistics of the CES-D score, insight, cause, positive and negative emotion words from the the e-mails of the clients on the pre- (T_0) and post-therapeutic (T_1) measurement.

Variable	Time	M	SD	Med.	Min.	Max.
CES-D	T_0	23.41	7.51	23	10	49
	T_1	15.42	8.07	14	1	37
positive emotion	T_0	36.78	20.73	35	2	110
	T_1	43.71	32.07	34	0	162
negative emotion	T_0	25.47	16.17	22	1	77
	T_1	17.76	13.76	14	0	62
insight	T_0	50.52	27.03	50	1	143
	T_1	45.86	29.84	41	2	173
cause	T_0	21.32	12.92	18	2	59
	T_1	20.54	15.25	18	0	85

Note. 'Med.' is an abbreviation for the median score of the variable; 'Min.' and 'Max.' refer to the minimum and maximum score.

two frameworks should be combined, without presenting results of statistical significance.

vi

Variable descriptions

In total, we used five variables, one from the intervention from Lamers et al. (2015); we obtained the other four (text) variables from the LIWC program by Pennebaker, Francis, et al. (2015). The CES-D score ($M = 19.42$, $SD = 8.75$), the number of positive emotion words ($M = 40.25$, $SD = 27.15$), the number of negative emotion words ($M = 21.62$, $SD = 11.73$), the number of insight words ($M = 48.19$, $SD = 28.49$), and the number of cause words ($M = 20.93$, $SD = 14.1$) are summarized in Table 6.1 (mean and standard deviations in the text are combinations of the pre- and post-therapeutic measurements).

Multilevel models

In total, we estimated five multilevel models (see Figure 6.1). The post-therapeutic measurement of CES-D was the main outcome. In Mo, model 0, we estimated the post-therapeutic measurement based on a random intercept for each client.

Mo is nested under the other four models. In ME1 ('Model *Emotion*'), we estimated the post-therapeutic effect of the number of positive emotions words and a random intercept for each client. ME2 was similar to ME1, but instead of positive emotion words, we estimated the effect of (the number of) negative emotions words. MCP1 ('Model *Cognitive Process*') was similar in the same respect: we estimated the effect of insight words (instead of positive or negative words), and in MCP2 we estimated the effect of cause words.

```
# Modelling the null / basic model.
M0 <- lmer(cesd ~ post + (1 | id), data)

# Modelling of the positive and negative words categories.
ME1 <- lmer(cesd ~ post*posemo + (1 | id), data)
ME2 <- lmer(cesd ~ post*negemo + (1 | id), data)

# Modelling of the insight and cause categories.
MCP1 <- lmer(cesd ~ post*insight + (1 | id), data)
MCP2 <- lmer(cesd ~ post*cause   + (1 | id), data)
```

Figure 6.1: `R` code of the five multi-level models (Mo, ME1, ME2, MCP1, and MCP2) using the `lme4`-package. In all models, we estimated the post-therapeutic measurement of CES-D (`cesd`) based on a random intercept for each client (`id`). In ME1 we estimated the post-therapeutic effect of the number of positive emotion words as the interaction effect between the number of positive emotion words (`posemo`) and an indicator variable (`post`). The other models have a similar structure but different variables: in ME2 we estimated the effect of the number of negative emotion words (`negemo`), in MCP1 we estimated the effect of the number of `insight` words, and in the MCP2 we estimated the effect of the `cause` words. Mo is nested under each of these models. In `R`, a hashtag (`#`) comments a line, instructing `R` to ignore whatever text comes after the hashtag.

Interpretation of the results

The data do not support our hypotheses that the writing intervention improves the number of positive, insight and cause words, while decreasing the number of negative words. Rather than using the data of Lamers et al. (2015) as a case to obtain new insights about TCPR, we present it as a use case for process researchers who wish to investigate e-mail data through multilevel models. Accordingly, we assessed the results in Table 6.2 in four steps.

Table 6.2: Model fit, parameter estimates and corresponding standard errors of the fixed and random effects of the five multilevel models.

	Baseline	Emotion		Cognitive Processes	
	M0	ME1	ME2	MCP1	MCP2
Fixed					
intercept	23.41 (0.79) *	22.83 (1.52) **	22.61 (1.40) **	21.62 (1.59) **	21.69 (1.45) **
post-indicator	-7.99 (0.86) *	0.02 (0.04)	0.03 (0.05)	0.04 (0.03)	0.08 (0.06)
variable †		-5.60 (1.74) *	-6.33 (1.65) **	-5.33 (1.82) *	-5.50 (1.66) **
interaction ‡		-0.06 (0.04)	-0.08 (0.07)	-0.05 (0.03)	-0.12 (0.07)
Random					
σ_e^2	35.74	35.37	35.80	35.46	35.66
τ	25.06	24.89	25.21	25.21	24.72
Model fit					
deviance	1327.38	1323.48	1325.89	1324.58	1324.24
AIC	1335.38	1335.48	1337.89	1336.58	1336.24
BIC	1348.45	1355.09	1357.50	1356.19	1355.85
logLik	-663.69	-661.74	-662.95	-662.29	-662.12
χ^2		3.89	1.48	2.80	3.13
χ^2 df		2	2	2	2
Effect size					
Ω_0^2 df	0.67	0.68	0.67	0.68	0.67

Note. Coefficients (and standard errors). * $p < .01$; ** $p < .001$

† The mean of the text-variable, indicated by 'variable' in the Table, changes between the five models: in ME1 it is the number of positive words, in ME2 it is the number of negative words, in MCP1 it is the number of insight words, and in MCP2 it is the number of cause words.

‡ The 'interaction' variable is the interaction between the text variable (see †) and the post-therapeutic indicator ('post-indicator').

1. Fixed effects: intercept and post-therapeutic indicator

The post-therapeutic effect of the writing intervention is estimated as the interaction ('interaction' in Table 6.2) between the model specific variable ('variable', with a varying meaning between the models, variable indicates the number of positive emotion words in ME1, negative emotion words in ME2, insight words in MCP1, and cause words in MCP2) and the post-therapeutic indicator in Table 6.2. As we are specifically interested in the post-therapeutic interaction effect, we do not interpret the effect of the model specific variable and post-therapeutic indicator in Table 6.2. The fixed effect of Mo is the grand mean (μ), which is interpretable as the positive effect of the writing treatment, without specific change effects of the word categories we included. We also estimated the effect of the post-therapeutic indicator. However, this effect should not be interpreted, as it merely acts as a dummy variable in our model.

2. Assess post-treatment effects

There are two ways to evaluate the model(s). The first is based on values of the post-therapeutic interactions. Table 6.2 does not give an indication that models ME1, ME2, MCP1 and MCP2 have significant post-treatment effects at the $(p < .05)$ level. Because all the relevant information lies in the interaction effect, the effect of the (text-)'variable' should also not be interpreted.

The second way to evaluate models is based on model fit. Of the all the model fit information in Table 6.2, the χ^2-test is perhaps the most straightforward to interpret, as it comes with a significance test. As none of the χ^2-tests are significant, the model fit information in Table 6.2 does not indicate that one of the four models (ME1, ME2, MCP1, and MCP2) is a (significant) improvement over the baseline model Mo. The other fit criteria should be seen as measures that indicate good model fit if they are closer to zero (there are several good sources, we suggest Burnham & Anderson, 2004, as a starting point).

3. Random effects

The variance of the random effect τ express the variation in post-therapeutic depression scores for individuals. The variance of the residual error σ_e^2 expresses the variance of the measurement errors, conditional on the individuals (the random effects). Table 6.2 shows that the main effect of the text variables are – relative to the interaction effects– quite large. This is an indication that the sample (and population) are quite heterogeneous, making it difficult to estimate the

effect of the writing intervention, as homogeneous treatment effect are simpler to estimate.

4. Effect size

For the calculation of the effect sizes, we followed the suggestions of Xu (2003): Ω_0^2 in Table 6.2 is a generalization of the well-known R^2 measure, which can be interpreted as a measure for explained variance in multilevel models. Overall, Table 6.2 shows that all models have a relative large proportion of explained variance. However, as model fit is (decimally) similar for all models, we cannot conclude that one model should be preferred over the others.

Discussion

Key questions of Therapeutic Change Process Research (TCPR) usually adhere to obtaining a thorough understanding of the change processes that are (most) beneficial to the client. For TPCR, the pertinent question is not whether psychotherapy is effective, but how change occurs. It is common for TCPR to study the language used in the (therapeutic) interaction between client and counsellor in order to obtain answers to this question. Two challenges arise, how to obtain text-measures that relate to change processes, and how to analyse these change processes. We argued that text mining could be used for the first challenge, and multi-levelled models (MLMs) to overcome the second.

Conclusion

The complete-data subset from Lamers et al. (2015) does not suggest that the writing intervention contributes to change in the (number of) insight, cause, positive and negative emotion words. The analyses show that the intervention does decrease post-therapeutic depression, however, the data did not indicate that this decrease could be associated with one of the text variables.

We aimed to make a case for the correct analyses of e-mail data, by obtaining text variables from large bodies of text, not to obtain theoretical insights. We showed that text mining is an appropriate tool to model change processes, as it can answer questions related to change processes.

The second goal of our paper was to show how complex and multi-layered change processes should be assessed. We presented a straightforward re-parametrization of multi-level models, that allowed for assessing post-therapeutic change. The way we parametrize our MLMs allows for modelling a baseline (pre-therapeutic

score) and change (post-therapeutic score) over time, while accounting for the dependency between pre- and post-therapeutic score of each client. This also corresponds to growth modelling of multilevel data, where measurements are nested within subjects (Muthén, 1997). The association of specific text variables to the outcomes of the intervention was illustrative for these two points. Based on this proof-of-concept, we conclude that obtaining and analyses of textual information through text mining and MLMs can indeed advance TCPR.

Relevance

The main advantage of these models is that it opens up the possibility to engage more with clients in therapeutical settings. With web-based interventions on the rise, there is clear room to do so. The information from texts, which is directly accessible and does not require intensive transcription procedures, and can then be used to steer the therapeutic process in the desirable direction. Text mining can thus be used as a form of 'direct feedback', as MLMs allow for correct modelling of the relations between variables.

Open challenges

We proposed that text mining can be used to identify the important change processes within therapy related texts, and MLMs can be used to explain the relations between processes and outcomes. Full demonstration of the capabilities of this framework requires multiple datasets, and many of the problems that we faced require the attention of more researchers. We start the discussion session by describing these (open) challenges. Then, in the next section, we cover the limitations specific to our study.

Operationalization

Operationalization is one of the first challenges that users of text mining for TCPR face. Many of the TCPR constructs are theoretical, and need to be operationalized into linguistic features so that they are clearly distinguishable, measurable, and understandable in terms of empirical observations. Examples of these variables include emotional ventilation, dramatic relief, tension release, abreaction, or catharsis (for more examples, see Grencavage & Norcross, 1990). Operationalization is not only an important aspect for TCPR, nor is it limited to psychology, the whole social and life sciences require good operationalizations.

The linguistic products of therapy (diaries, psychotherapeutic assignments, or transcripts of the therapeutic interaction) provide rich source of research ma-

terial, provided that the variables of interest are adjustable to texts. In our current work, we used a basic text features from LIWC. We justified our use of these basic text features because we aimed to give a proof-of-concept with the intend of showing how TCPR and MLM can be bridged.

However, our choice for such a basic text variable leaves one of the largest challenges open: what to (text) mine? Traditionally, the text mining community was more concerned with collecting, storing and managing large bodies of unstructured text rather than applying theoretical models from other fields. Advances in the field of computer science made technical issues less insurmountable than they were a decade ago (Mayer-Schönberger & Cukier, 2013, p. 8). As a results, text mining is no longer reserved for those with a computer science degree.

The increase in solved technical issues did not lead to insights in 'what to mine'. We did not aim to advance TCPR theory with our current paper; we intended our work as a method paper, because with the current state of the literature, it is difficult to determine which textual predictors are meaningful. Also, we feel that our proposition to bridge text mining and MLMs itself allows for advancing TCPR theory. Constructs as described by Elliott (2010, 2012), Grencavage and Norcross (1990), and Orlinsky et al. (2004) require a 'translation', or adjustment, before text mining is applicable to these data types. Domain experts in the TCPR field are well-equipped to face this question, but this requires an interdisciplinary approach.

We showed how MLMs and text mining can be combined, but our proposition leaves open how TCPR concepts should be operationalized for text mining metrics. That would require an interdisciplinary collaboration and discussion. However, the future does look bright: based on our proof-of-concept study we conclude that MLMs and text mining can indeed advance TCPR.

The next step in that direction, would be to –aside from LIWC– incorporate other existing text mining software, such as TCM (*Therapeutic Cycles Model*; Mergenthaler, 1996), or CALAS (*Computer Assisted Language Analysis System*; Anderson et al., 1999).

Measurement error

Elliott (2010) argued that TCPR is plagued by measurement error. Although the term 'error' is often used, in our experience, it can refer to two different concepts depending on the field of study. With the risk of over-generalization, in the machine learning community and other fields that rely heavily on predictive analytics, error often refers to the error or confusion matrix. The table of confusion

reports the number of false positives and negatives, and the true positives, and negatives. These measurement represent the performance of an algorithm. Error then refers to measures of predictive error, the difference between the observed values and the values predicted by the model.

In statistics, error is related to measurement error, which represents the difference between a measured value of a quantity and its true value. Measurement error is often used to indicate whether or not measurement is reliable. Reliability expresses how repeatable measurements are when remeasured. The reliability of a measure is then a direct function of the amount of error is present in the measurement. Because no behavioural measure is perfectly reliable, some degree of measurement error will always occur. Therefore, reliability is low when there is a abundance of error, and vice versa. The underlying idea is that every observation is a combination of the hypothetical true score plus some measurement error.

Although nowadays ideas appear to be floating freely between machine learning and statistics (Wasserman, 2010, p. 8), some concepts –such as measurement error– are traditionally more associated with one branch rather than the other (see for example Donoho, 2017). Measurement error is well-established in statistics, and has potential for machine learning disciplines such as text mining. Variables are simply an operationalization of the process, behaviour or item that we are trying to measure. Estimation of the measurement error reflects the uncertainty present in the estimate. Consistency of the research measures benefits when accounting for measurement error.

In fact, with respect to measurement error, MLMs are the way forward. MLMs recognize the existence of several levels, nesting and hierarchies in data. MLMs capitalize on this concept by allowing for the inclusion of residual components at each level of the hierarchy. Hence, the precision of the estimation of measurement error increases, as the residual variance is partitioned.

Sample size

TCPR is rooted in qualitative research methods; MLMs come from the quantitative sciences. Intensive case-studies are not uncommon for qualitative scientists, but will lead to statistical power issues for MLMs. As MLMs introduce multiple levels, the total number of units observed for each level become the sample size. The relevant sample size for power issues depends on the parameters that are being tested. Unlike the traditional regression, there is a difference between testing a regression coefficient or a variance parameters in a MLM.

The main limit is the sample size at the highest level of organization. Natu-

rally, having multiple measures (at the first level) for one client (second level) is less informative then having these same multiple measures for multiple clients. The number of clients will therefore be one of the main issues for using MLMs for TCPR, but it will limit the wide scale application of MLMs for TCPR.

Limitations

We already gave an impression of some overarching open challenges that –in their current form– limit the applicability and wide-scale impact of the ideas we presented in the current work.

Excluded therapy

Based on the design of Lamers et al. (2015), it would also have been straightforward to model the effect of treatment. Modelling treatment as a random effect could have provided an insight in the efficacy of the treatment for each individual client. The fixed effect of treatment would have given some insight in the average efficacy of the treatment groups in comparison with each other.

We however, as Lamers et al. (2015), we could not differentiate between the two conditions of the treatment. They found both writing conditions to be helpful in comparison to the control group, but could not differentiate between the expressive writing and autobiographical writing conditions.

We justified our exclusion furthermore because we only intended to show that text mining can be used to obtain additional predictors for multilevel models. Our intend was not to offer new theoretical insights for psychological writing interventions; we intended to offer methodological rather theoretical insights.

Complete cases

We only included clients with complete cases and did not attempt to account for the missing data. First of all, it was difficult to determine why certain measurements where missing for an individual. Lamers et al. (2015) gave an overview of drop-out and missing data: it was challenging for us to determine post-hoc what the exact reason for missing data or drop out was for an individual based on general information.

Because we did not understand the underlying reason for the occurrence of missing data, we were hesitant in choosing an imputation technique. Also, because we did not intend to draw theoretical conclusions from our work, we felt

that the issues with generalisation and validity associated with ignoring missing data were less relevant for our proof-of-concept.

Future research

MLMs come with the well-known advantage that the model can incorporate the hierarchical structure of the data. This is idea holds potential for TCPR, as change processes are often multifaceted and multi-layered. For example, an interesting analyses would be to see the effect counsellors have on their clients. As a counsellor almost always treats multiple clients, it is possible to estimate the effect of a counsellor on its clients. Combining this form of nesting with other forms of nesting, such as the treatment effect itself, it would then be possible to estimate counsellor efficacy in different arms of the treatment. Accounting for clustering influences the estimation of the treatment effect as these influences are expressed as parameters in the model.

TCPR would also receive an enormous boost when change processes could be automatically detected through text mining. Some methods, such as the *Innovative Moments Coding Scale* (perhaps better known under its abbreviated name ICMS, see Gonçalves et al., 2009; Gonçalves et al., 2010), already provided an avenue for doing so.

We are optimistic about TCPR's future through the happy marriage between text mining and MLMs. Especially in the social sciences, many phenomena can considered to be levelled, and the usage of text mining is already picking up. Social scientists in general often intend to learn about relations between variables in the population. In our view, in comparison with machine learning models, MLMs are of use to social scientists because they can provide theoretical insights in the relationships between, rather than building a black box model with the goal of attaining good predictive qualitative. MLMs can thus be used to explain relations between variables, whereas text mining can thus be used to obtain important therapy related variables, given that other TCPR research point in the direction of which important constructs are present in texts.

What
Works
When
for
Whom

Part III

An alternative TCPR model

What
Works
When
for
Whom

In the social sciences, data structures are often hierarchical in the following sense: We have variables describing individuals, but the individuals also are grouped into larger units, each unit consisting of a number of individuals. We also have variables describing these higher order units.

Once we have discovered this one example of a hierarchical data structure, we see many of them.

Moreover, many repeated measurements are hierarchical. If we follow individuals over time, then the measurements for any particular individual are a group, in the same way as the school class is a group. If each interviewer interviews a group of interviewees, then the interviewers are the higher level. **Thinking about these hierarchical structures a bit longer inevitably leads to the conclusion that many, if not most, social science data have this nested or hierarchical structure**.

Jan de Leeuw (Raudenbush & Bryk, 2002b, p. xix, xx)

Process research should be conducted at multiple levels of abstraction.

However, **additional refinement of statistical methods is required to fully account for the multilayered complexity of therapeutic processes**.

Knobloch-Fedders et al. (2015)

What
Works
When
for
Whom

7 Small and Negative Correlations Among Clustered Observations: Limitations of the Linear Mixed Effects Model

Abstract

The linear mixed effects model is an often used tool for the analysis of multilevel data. However, this model has an ill-understood shortcoming: it assumes that observations within clusters are always positively correlated. This assumption is not always true: individuals competing in a cluster for scarce resources are negatively correlated. Random effects in a mixed effects model can model a positive correlation among clustered observations but not a negative correlation. As negative clustering effects are largely unknown to the sheer majority of the research community, we conducted a simulation study to detail the bias that occurs when analysing negative clustering effects with the linear mixed effects model. We also demonstrate that ignoring a small negative correlation leads to deflated Type-I errors, invalid standard errors and confidence intervals in regression analysis. When negative clustering effects are ignored, mixed effects models incorrectly assume that observations are independently distributed. We highlight the importance of understanding these phenomena through analysis of the data from Lamers et al. (2015). We conclude with a reflection on well-known multilevel modelling rules when dealing with negative dependencies in a cluster: negative clustering effects can, do and will occur and these effects cannot be ignored.

Keywords: negative clustering effects, negative cluster correlation, negative ICC, covariance structure models, Linear Mixed Effects model

Nielsen, N. M., Smink, W. A. C. & Fox, J. P. *In press at Behaviormetrika*.

Introduction

The popularity of the linear mixed effects models (e.g., random effect models, multilevel models) is intuitively explained by the variety of different names under which the family of statistical models for clustered data are known. In clustered data (e.g., hierarchical data, multilevel data) observations are associated, and not independently observed (Dorman, 2008). Dependencies among observations in clusters can be expressed as a correlation, where positively correlated observations share similar information (Kenny & Judd, 1986), and are not as informative as independent observations (Galbraith et al., 2010). Only when accounting for the correlation between clustered observations, correct statistical inferences can be made. In the well-known multilevel modelling framework (McCulloch et al., 2008), this correlation is modelled by a random effect –also known as a latent variable– where clustered observations are positively correlated since they share the same random effect. The variance of the random effect then determines the strength of the correlation. As a result, the multilevel modelling frameworks restricts correlations to be positive, since a variance parameter cannot be negative.

However, negative correlations among clustered observations *can* and *do* occur (Kenny et al., 2002). For instance, when fixed resources are divided among group members, in non-random sampling when dissimilar groups are sampled by intention, or when there is competitive social interaction: when individuals compete for a scarce (and fixed) set of resources (e.g. litter mates are negatively correlated in terms of food, water and living space), and the speaking time of one individual is at the expense of another individual (Pryseley et al., 2011). When observations within clusters are negatively associated, observations within clusters are less alike than observations from different clusters (Kenny & Judd, 1986). From a sampling perspective this is sometimes referred to as the situation where observations within a cluster are even less alike than under random assignment of observations to clusters (Molenberghs & Verbeke, 2007, 2011; Verbeke & Molenberghs, 2003). Negative intra-cluster correlations (ICC; see Table 7.1 for an overview of the often used abbreviations) can also be detected in randomized experiments, when evaluating the effects of covariates that vary systematically within each cluster (Norton et al., 1996).

In general, it is well-known that ignoring a small positive clustering leads to the incorrect assumption that the observations are independently distributed. It is our aim to extend this knowledge with the current article: we will show that ignoring a positive *and negative* clustering leads to a violation of the independence assumption. In fact, any violation of the independence assumption

Table 7.1: List with the often used abbreviations in the current article.

Abbreviation	*Full term*
CI	Confidence Interval
CR	Coverage Rate
ICC	Intra-Cluster Correlation, Intra-Class Correlation, Intra-class Correlation Coefficient
LM	Linear Model
LME	Linear Mixed Effects (model)
CSM	Covariance Structure Model
SE	Standard Error
VIF	Variance Inflation Factor, also known as the design effect

(positive and negative) results in inaccurate Type-I errors, which increase the risk of accepting an incorrect hypothesis (P. Clarke, 2008). Barcikowski (1981) quantified the effects of ignoring small positive correlations in clustered observations in a two-level study design (with a group and an individual level). He showed that, when having ten observations per group, even the ignorance of an ICC as small as .01 can lead to an inflation of making a Type-I error: a regression effect will be assumed to be significant with a significance level of 5% although the true significance level equalled 6%. Furthermore, Barcikowski showed that the Type-I error increased for increasing values of the ICC. For an ICC of .05 the Type-I error rate is .11 and for an ICC of .40 the Type-I error is .46. Moreover, by increasing the number of observations per group the Type-I error is even more inflated (the findings of Barcikowski are in line with those of many others, see for example P. Clarke, 2008; Dorman, 2008; Rosner & Grove, 1999).

As negative clustering effects are largely unknown to the sheer majority of the research community, we conducted a simulation study to detail the bias that occurs when analysing negative clustering effects with the linear mixed effects model in similar fashion as Barcikowski (1981). Towards that end, we demonstrate that ignoring a small negative correlation leads to deflated Type-I errors, invalid standard errors and confidence intervals in regression analysis. We highlight the importance of understanding these phenomena through analysis of the data from Lamers et al. (2015). We conclude with an updated reflection on well-known multilevel modelling rules. In the remainder of this section, we discuss negative dependencies between observations in clustered data, show how the Linear Mixed Effects model (LME) deals with negative clustering effects, and reflect on why the LME should include negative variance components.

Type-I errors and positive and negative dependencies between observations

The inflation of the Type-I error under violated of the independence assumption can be explained by the variance inflation factor (VIF), also referred to as the design effect (Kish, 1965). In case of cluster sampling, a design effect that is greater than one is known to indicate a positive within-cluster correlation, indicating that observations are not independent of each other. When $VIF > 1$ the precision of cluster sample estimates are less than that of those based on a simple random sample with a similar size. The homogeneity in clustered observations leads to less information in comparison to an independent random sample. When ignoring a small positive ICC, the VIF is underestimated, which leads to an underestimation of the standard errors (i.e. overestimating the precision), and the corresponding confidence intervals (CIs) are too narrow, and effect sizes will then also be incorrect as they depend on standard error (SE) estimates (Hox et al., 2010; Kenny et al., 2002). When the CI of an estimate is too narrow, there is an increase in the probability to reject a correct null hypothesis, which corresponds to an inflation of the Type-I error.

Although a few researchers have reported about negative clustering effects (El Leithy et al., 2016; Kenny et al., 2002; Klotzke & Fox, 2019a, 2019b; Loeys & Molenberghs, 2013; Molenberghs & Verbeke, 2007, 2011; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003), the effects of ignoring negatively clustered observations has hardly been recognized. Because negative clustering effects are not considered by the majority of the multilevel modelling community, these effects are not well understood. This is partly caused by the fact that the mixed effect models (to which we also refer as 'LME', see Table 7.1) can only describe positive correlations, and cannot handle negative correlations among clustered observations (Searle et al., 1992). In the next section, it is explained *why* negative clustering effects cannot be modeled with LME, and we reflect on the key principles of negative ICCs.

The LME cannot identify any negative correlation and will assume independently distributed observations. Researchers usually fix negative ICC estimates to zero and ignore any negative correlation within a cluster (Baldwin et al., 2008; Maas & Hox, 2005). Furthermore, it is sometimes concluded that negative ICC estimates are caused by a small between-cluster variance (smaller than the within-cluster variance) and that such a small between-group variance can be ignored (Giberson et al., 2005; Krannitz et al., 2015; Langfred, 2007). Other researchers relate negative ICC estimates to sampling error (cf. Eldridge et al., 2009), which can be ignored. Others –such as Baldwin et al. (2008) and Norton

et al. (1996), and Rosner and Grove (1999)– stated that the Type-I error will be deflated when fixing a negative ICC to zero.

The Linear Mixed Effects model and negative dependencies

In this study, we consider two models: the LME and a Covariance Structure Model (CSM, see Table 7.1). Both models can assess clustered data, where a one-way classification structured is considered. In the one-way classification, a common correlation is assumed among clustered observations, and observations from different clusters are assumed to be independently distributed.

The Linear Mixed Effects model

Without making an explicit distinction between a random variable and a realized value, the LME for the one-way classification is given by

$$y_{ij} \quad = \quad \beta_0 + \beta_1 X_{ij} + u_j + e_{ij}, \tag{7.1}$$

where the random effect is assumed to be normally distributed, $u_j \sim \mathcal{N}(0, \tau)$, and the error term is also assumed to follow a normal distribution $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. A total of $j = 1, \ldots, m$ clusters are assumed with each $i = 1, \ldots, n$ observations, which leads to a balanced study design. The common intercept and regression parameter are referred to as β_0 and β_1, respectively. The outcome y_{ij} is assumed to be independently distributed given the random effect u_j.

It can be shown that the random effect u_j defines a variance-covariance structure for the data. The covariance between two clustered observations is equal to

$$\begin{aligned} cov\left(y_{ij}, y_{lj}\right) &= cov\left(E\left(y_{ij} \mid u_j\right), E\left(y_{lj} \mid u_j\right)\right) + E\left(cov\left(y_{ij}, y_{lj} \mid u_j\right)\right) \\ &= cov\left(\beta_0 + \beta_1 X_{ij} + u_j, \beta_0 + \beta_1 X_{lj} + u_j\right) + 0 \\ &= cov\left(u_j, u_j\right) = var\left(u_j\right) = \tau, \end{aligned} \tag{7.2}$$

and the variance of an observation equals

$$\begin{aligned} var\left(y_{ij}\right) &= var\left(E\left(y_{ij} \mid e_{ij}\right)\right) + E\left(var\left(y_{ij} \mid u_j\right)\right) \\ &= \sigma^2 + \tau. \end{aligned} \tag{7.3}$$

The dependence structure of the observations in the clusters $\mathbf{y}_j$ modelled by the random effect u_j is given by

$$var\left(\mathbf{y}_j\right) = \mathbf{\Sigma} = \begin{bmatrix} \sigma^2 + \tau & \tau & \dots & \tau \\ \tau & \sigma^2 + \tau & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau & \dots & \tau & \sigma^2 + \tau \end{bmatrix}. \tag{7.4}$$

Thus, $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n + \mathbf{J}_n\tau$ represents the dependence structure implied by the random effect u_j. The $\mathbf{J}_n$ is a matrix of dimension n with all elements equal to one and $\mathbf{I}_n$ is the identity matrix.

The Covariance Structure Model

An alternative specification of the LME in Equation (7.1) can be given. In this approach, the covariance structure is modelled directly and not indirectly through the specification of a random effect. The distribution of clustered observations is assumed to be multivariate normally distributed with a covariance matrix $\mathbf{\Sigma}$,

$$\mathbf{y}_j \quad = \quad \beta_0 + \beta_1\mathbf{X}_j + \mathbf{e}_j, \tag{7.5}$$

where the errors are multivariate normally distributed, $\mathbf{e}_j \sim \mathcal{N}(0, \mathbf{\Sigma})$. We refer to the model in Equation (7.5) as the CSM. The development and use of the covariance structure model has a long history, which is intertwined with the development of factor models. Classic works in covariance structure modelling can be found in that tradition (e.g., Bock & Bargmann, 1966; Jöreskog, 1969, 1971). Fox et al. (2017) and Klotzke and Fox (2019a), and Klotzke and Fox (2019b) developed a novel Bayesian modelling framework in which they directly modelled the covariance structure of more complex dependence structures. In their *Bayesian Covariance Structure Modelling* (BCSM) approach, dependencies among observations that are usually modelled through random effects are modelled directly through covariance parameters under the BCSM.

When comparing the modelling structure of the CSM (also referred to as BCSM; Klotzke & Fox, 2019a, 2019b) with that of the LME, it can be seen that the τ is restricted to be positive in the model in Equation (7.1), since it represents a *variance* parameter. However, in the model in Equation (7.5), the τ parameter can also be negative since it represents a *covariance* parameter. This makes the CSM more general than the LME, since the covariance parameters can be positive and negative, which allows for more flexibility in specifying complex dependence

structures (cf. Klotzke & Fox, 2019a, 2019b).

The Linear Mixed Effects model with negative variance components

There are some restrictions on the variance-covariance components in the CSM. From the definition of the error variance follows directly that the σ^2 is restricted to be greater than zero (i.e. $0 < \sigma^2 < \infty$). However, the covariance parameter τ is not necessarily restricted to be greater than zero. Under the CSM, the covariance matrix $\mathbf{\Sigma}$ needs to be positive definite, which that implies the restriction –for balanced designs– $n\tau + \sigma^2 > 0$. This important result follows from Rao (1973, p. 32), where the determinant of a compound symmetry covariance matrix is expressed as

$$\begin{aligned} det\left(\sigma^2\mathbf{I}_n + \tau\mathbf{J}_n\right) &= det\left(\sigma^2\mathbf{I}_n\right)\left(1 + \tau\mathbf{1}_n^t\mathbf{1}_n/\sigma^2\right) \\ &= \sigma^2\left(1 + n\tau/\sigma^2\right) = n\tau + \sigma^2, \end{aligned} \tag{7.6}$$

and the covariance matrix is positive definite if the determinant is greater than zero . Subsequently, τ needs to be greater than $-\sigma^2/n$. However, when modeling the covariance structure with the LME, the τ is restricted to be greater than zero, since it represents the random intercept variance. In the literature, it has been shown that the maximum likelihood estimate of the random effect variance can become negative (El Leithy et al., 2016; Kenny et al., 2002; Klotzke & Fox, 2019a, 2019b; Loeys & Molenberghs, 2013; Molenberghs & Verbeke, 2007, 2011; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003). For the (one-way) LME (for balanced groups), two sums of squares are considered to estimate the covariance components τ and σ^2,

$$\begin{aligned} SS_A &= \sum_{j=1}^{m} n\left(\overline{y}_j - \overline{y}\right)^2, \\ SS_E &= \sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij} - \overline{y}_j\right)^2. \end{aligned} \tag{7.7}$$

Consider the sum of squares SS_A, which has as expected value $n\tau+\sigma^2$. It follows that, $\hat{\tau} = SS_A/(nm)-\sigma^2/n$, which leads to a negative estimate of τ if $\sigma^2 > SS_A/m$. This scenario is often neglected or referred to as statistically incorrect, restricting τ to represent a positive covariance among clustered observations.

For $\tau > 0$, the ICC is often interpreted as the ratio of variance explained by

the clustering of observations in comparison to the total variance in the data; $\rho = \tau/(\tau + \sigma^2)$ (Oliveira et al., 2017; Raudenbush & Bryk, 2002b; Snijders & Bosker, 2012). However, the ICC can also be considered to quantify the degree of resemblance or average similarity of observations within a cluster, or as the 'average correlation' in each cluster (Kenny & Judd, 1986; Kenny et al., 2002). Then, conceptually, a negative covariance ($\tau < 0$) represents a negative ICC. In that case, ρ becomes negative, and the ICC represents a negative association among clustered observations (i.e. observations within clusters are less alike than observations from different clusters). A negative ICC simply represents the opposite of a positive ICC: if an observation in a cluster is below the population mean, then it is more likely that another value in that cluster is above the population mean if the observations are negatively correlated (Kenny & Judd, 1986; Kenny et al., 2002).

Even though negative clustering effects were discussed previously by others (cf. Oliveira et al., 2017; Pryseley et al., 2011), there still appears to be a lack of awareness about these effects. As the LME comes with the restriction that observations need to be positively clustered, several suggestions can be found in the literature that τ should be set to zero, when the ICC estimate becomes negative (see for example Baldwin et al., 2008; Eldridge et al., 2009; Gibson et al., 2015; Krannitz et al., 2015; Langfred, 2007; Maas & Hox, 2005). In the next section, we will discuss our simulation study which aims to not only show that fixing the ICC to 0 is –in fact– wrong, but also quantify the bias that arises when negative clustering effects are ignored.

Method

The object of the simulation study was to quantify the estimation errors that are made in the statistical analysis, when ignoring a (small) negative or positive clustering in the data. While ignoring a small positive and negative ICC, the accuracy of the intercept and regression parameter estimates, the SEs, the 95% CIs, and the inflation and/or deflation of Type-I errors were assessed.

A large number of clusters can compensate for biases that occur due to ignoring small ICC values, where the number of observations per cluster affects the magnitude of the Type-I error (Barcikowski, 1981; Dorman, 2008). In this study, $m = 10$ clusters were considered to assess the bias, and the number of observations per cluster $n = \{10{,}15{,}30\}$ varied to examine the effects for small sample sizes. The error variance σ^2 was fixed to one, where τ took on values ranging from $-\sigma^2/n$ to 0, in incremental steps of .02 and from 0 to .20 in steps

of .05. The step size for $\tau < 0$ were set to .005 for $n = 30$, so that at least two negative values of τ were used for data generation. For each condition a lower-bound was defined, $\tau_{Lb} = -\sigma^2/n$, which represented the lower bound on the allowable negative correlations in the clustered data. Only the CSM allowed generating data with negative correlations, where under the LME data could only be generated with positive correlations. The true value of the intercept and slope parameter were set to 0 and 0.1 ($\beta_0 = 0, \beta_1 = 0.1$), respectively, and those parameters were considered to assess the effects of ignoring the correlation in the data. An intercept-only model and an intercept-slope were used to simulate data.

A Monte Carlo simulation study was used to evaluate the appropriateness of the LME as an analysis tool for negative clustering and small positive clustering effects. Therefore, data were generated under the LME (see Equation (7.1)) which only generated data with positively correlated observations in clusters. Data with negatively or positively correlated observations were generated according to the CSM (Equation 7.5), with a covariance matrix displaying the dependence structure in a cluster. The LME was fitted to data generated under the LME and under the CSM, and the parameter estimates (REML) for the LME were obtained using the `lme4`-package in `R` (D. Bates et al., 2015; R Core Team, 2020). A linear regression model (LM) was also fitted that ignored any correlation in the clustered observations (positive or negative). Parameter estimates for the LM were obtained using the `lm`-function in the `stats`-package in `R` (the `stats`-package is a core package in R Core Team, 2020). The `lmerTest`-package was used to compute p-values of the fixed effects in the LME (Kuznetsova et al., 2017), where Satterthwaite's degrees of freedom method was used (Satterthwaite, 1946). For each condition, a total of 1,000 data sets were generated, and reported per condition the average p-value, coverage rate (CR), SE, and bias of τ estimate. The bias was computed as the average over replications between the estimate and the true parameter value. The average SE estimate across replications was computed and reported as the SE. The CR was computed as the percentage of times the true parameter value was located in the 95% confidence interval over data replications. The average of the computed p-values across data replications was considered to be the average p-value. The ICC values were close to the τ values, since the error variance σ^2 was set to 1. Therefore, results were mostly reported for τ and only sparsely for the ICC.

Results

We use this section to first discuss the simulation study. We also include a real data example to illustrate that not accounting for negative clustering can lead to an increase of a Type-II error. For this example, we re-used data from Lamers et al. (2015), and Smink, Fox, et al. (2019).[1]

Coverage rate and Type-I error

The 95% CRs were estimated for the intercept-only condition ($\beta_0 = 0$, $\beta_1 = 0$), and the intercept-slope condition ($\beta_0 = 0$, $\beta_1 = .1$; see Equation 7.1). For the intercept-only condition, the results concerning the intercept are presented under the label β_0. For the intercept-slope condition, the results concerning the slope are presented under the label β_1. In Table 7.2, results are given of data generated under the LME, which were analysed with the LME (model variant 1 in Table 7.2) and the LM (model variant 3), and results are given of data generated under the CSM which were analysed with the LME (model variant 2) and the LM (model variant 4). For data generated under CSM, (model 2 and 4) true values of τ were also allowed to be negative. For data generated under the LME (model 1 and 3), the τ was restricted to be greater than or equal to zero. Note that when $\tau = 0$, the u_j were equal to zero representing no variance across clusters (see Equation 7.1). The simulated data for $\tau = 0$ under the LME did not contain any clustering effects. For data generated under the LME, the CR was only computed for $\tau \geq 0$.

The reported CRs of the intercept in the intercept-only condition showed bias. When the true value of τ was smaller than zero, the CR was greater than .95, which means that the Type-I error was deflated. For all cluster sizes, the CR was at least .97 for very small negative ICCs ($\tau = .01$), with a maximum of one for more negative ICCs. A similar bias in CR estimates were found for the LME and the LM (model variant 2 and 4 in Table 7.2). Thus, for negative correlations the LME performed similar to the LM. Under the LME, the random effect variance τ was estimated to be (approximately) zero, which made the LME similar to the LM as model for analysis for $\tau \leq 0$. Note that a CR of one is an upper bound, but the width of the CIs still increased for more negative values of τ.

When the correlation was greater than zero, and the LM was used to analyse the data, the CRs were highly underestimated, where the underestimation was larger for $n = 30$ than for $n = 10$. When increasing the correlation, the CRs were more underestimated. LM ignored correlation within groups, and the model

[1] Smink, Fox, et al. (2019) is chapter 6 of this thesis.

Table 7.2: 95% coverage rates for the intercept (β_0) and the slope parameter (β_1) for different cluster sizes ($n = 10, 15, 30$) and various correlations among clustered observations (τ).

	(1)		(2)		(3)		(4)	
GEN	LME		CSM		LME		CSM	
ANA	LME		LME		LM		LME	
τ	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	$L_b = -1/10, n = 10$							
−0.09			1.00	.93			1.00	.94
−0.07			1.00	.94			1.00	.94
−0.05			1.00	.95			1.00	.95
−0.03			.98	.95			.98	.95
−0.01			.98	.95			.97	.96
0.00	.96	.94	.97	.95	.95	.94	.95	.95
0.05	.93	.94	.94	.94	.89	.94	.90	.95
0.10	.94	.95	.94	.95	.86	.96	.86	.95
0.15	.94	.95	.93	.94	.81	.95	.80	.95
0.20	.94	.94	.94	.94	.76	.95	.80	.95
	$L_b = -1/15, n = 15$							
−0.06			1.00	.96			1.00	.96
−0.04			1.00	.95			1.00	.96
−0.02			.98	.94			.98	.94
0.00	.95	.95	.96	.94	.94	.95	.95	.94
0.05	.95	.95	.94	.95	.88	.95	.88	.95
0.10	.93	.96	.93	.95	.81	.97	.80	.95
0.15	.93	.94	.92	.95	.75	.94	.76	.96
0.20	.91	.94	.93	.96	.70	.94	.72	.95
	$L_b = -1/30, n = 30$							
−0.03			1.00	.95			1.00	.96
−0.01			.97	.96			.97	.97
0.00	.97	.95	.96	.94	.96	.94	.95	.94
0.05	.94	.96	.94	.95	.82	.95	.79	.94
0.10	.93	.94	.93	.93	.69	.93	.68	.93
0.15	.92	.95	.94	.94	.62	.94	.62	.92
0.20	.90	.96	.93	.95	.53	.94	.59	.94

Note. The τ is restricted to a lower bound of $-\sigma^2/n$. For data generated under the LME $\tau \geq 0$. LME is the linear mixed effects model, CSM is the covariance structure model, LM is the linear regression model. GEN, the model that *generated* the data; ANA, the model that *analysed* the data.

assumed more information in the data than there actually was observed. This led to an inflation of the Type-I error, since a significant result was more easily obtained with estimated CIs that were too narrow. This bias was not expected when the LME was used as model for analysis. However, in that case also an underestimation of the CRs was observed. For increasing value of the correlation a decrease in the CR was observed. This underestimation increased when increasing the cluster size, and was observed for data generated under CSM as well as for data generated under the LME. For medium cluster sizes ($m = \{15,30\}$), and relatively high correlations within a cluster ($.10-.20$), the CRs were underestimated, although the LME was used as the model for analysis. For all cluster sizes used in this simulation study, the coverage was at least .97 for small negative ICCs ($\tau \leq -.01$), which means that the Type-I error was deflated. This effect was the same for the CSM variants. The LME cannot handle negative cluster correlation. When the true τ was negative, τ was estimated to be around zero under the LME, leading to similar results of the LME and LM.

For all model variants and all conditions, the CR of the slope parameter was (approximately) equal to 95%, representing the finding that the 95% CI covered the true value in 95% of the data sets. The estimated CRs for the slope parameter did not show any bias. Even for negative correlations among clustered observations were the CRs around the level of 95%. It was concluded that the ignorance of the correlation within groups did not have an influence on the Type-I error of the slope parameter. Adjusting the level of correlation in the observations did not affect the regression effect or its precision of the predictor variable. As the slope was unaffected by the ignorance of the correlation, the remainder of this results sections only concerns data generated under the only-intercept condition.

Standard error and p-value

Generally, when testing if the intercept is equal to zero, $\beta_0 = 0$ and the true value is zero, then p-values are expected to be uniformly distributed and centred at .50. Furthermore, 5% of the p-values are expected to be smaller than or equal to .05. When the p-value is biased upwards, less than 5% of the p-values take a value of at most .05, which means that the Type-I error is deflated (i.e., a decrease in the probability to reject a correct null hypothesis). For p-values that are biased downwards, more than 5% of the p-values take a value of at most .05, which means that the Type-I error is inflated (i.e., an increase in the probability to reject a correct null hypothesis).

As can be seen in Table 7.3, for $\tau > 0$ and the LME or CSM was used to generate the data, the p-values were centred around .50 for different positive values of

τ. This confirmed that LME can be used to control for the positive correlation in clustered data, and correct p-values were computed for the intercept. When the LM was used to analyse the data, for $\tau > 0$, the p-values were underestimated. With increasing τ that was ignored, the model assumed more and more information in the data than there actually was, which led to a higher percentage of significant p-values than the significance level of 5%. This led to an increase of the probability of rejecting a correct null hypothesis using a significance level of 5%.

For τ equal to zero, the p-value was always biased upwards for data analysed with the LME, since estimates of τ were upwardly biased when the τ was negative or close to 0, see Table 7.4. This was caused by the lower bound for τ under the LME. For negative τ, the p-values were always overestimated, when analysing the data with LME or LM. This bias was similar under both models, since the LME and LM cannot control for negative cluster correlation. It can also be seen that the level of overestimation increases quickly for a small increase in negative τ.

The bias of the p-value was larger for larger cluster sizes, which was caused by a smaller sampling error and more data information about the intercept. For a large number of observations per cluster, already a small negative correlation led to a large bias in p-value. For example, for $n = 30$ the correct p-value of .50 was increased by .06 for $\tau = -.01$; and for $\tau = -.03$ the p-value was overestimated by .27. For smaller cluster sizes the bias was smaller. However, for a smaller cluster size the lower bound for τ was also smaller, and for more negative correlations among clustered observations the bias in p-values again increased.

Similar findings were made for the SE of the estimated intercept. When data was generated under the LME or CSM, for τ greater than zero, results under the LME showed that the SE estimates were slowly increasing for increasing values of τ. The increasing correlation in the data led to an increase in the reduction of information about the intercept. When the LM was used as model for analysis, the SEs were underestimated, since the correlation in the data was ignored. Both models cannot handle a negative correlation, which was ignored and the SE estimates showed a slight decrease for more negative correlations. The total variance in the observations was reduced for a negative τ (see Equation 8.3). This reduction in the total variance led to a reduction in the estimated SEs under LME and LM.

In Figure 7.1, it can be seen that for $\tau \geq .05$, correct estimates of the p-values (around .50) were obtained under the LME. For $0 \leq \tau < .05$, the correlation was (slightly) overestimated leading to an overestimation of the p-values under the LME. When accounting incorrectly for intra-cluster correlation the data con-

Table 7.3: Estimated SEs and p-values of the intercept (averaged across 1,000 replications) – which is set to 0 – for all four model variants, with varying number of observations per cluster (n).

	(1)	(2)	(3)	(4)
GEN	LME	CSM	LME	CSM
ANA	LME	LME	LM	LM
τ	SE (p)	SE (p)	SE (p)	SE (p)
	$L_b = -1/10, n = 10$			
−0.09		0.095 (.80)		0.095 (.80)
−0.07		0.097 (.68)		0.097 (.68)
−0.05		0.098 (.61)		0.098 (.61)
−0.03		0.100 (.57)		0.098 (.56)
−0.01		0.105 (.53)		0.099 (.51)
0.00	0.107 (.52)	0.107 (.53)	0.100 (.50)	0.100 (.50)
0.05	0.123 (.50)	0.124 (.52)	0.102 (.43)	0.102 (.45)
0.10	0.138 (.50)	0.139 (.50)	0.104 (.41)	0.104 (.40)
0.15	0.157 (.51)	0.156 (.50)	0.107 (.38)	0.106 (.37)
0.20	0.170 (.49)	0.168 (.51)	0.108 (.35)	0.108 (.36)
	$L_b = -1/15, n = 15$			
−0.06		0.079 (.76)		0.079 (.76)
−0.04		0.080 (.62)		0.080 (.62)
−0.02		0.083 (.57)		0.081 (.56)
0.00	0.087 (.52)	0.087 (.52)	0.081 (.49)	0.081 (.50)
0.05	0.108 (.52)	0.106 (.51)	0.084 (.43)	0.083 (.42)
0.10	0.127 (.50)	0.126 (.49)	0.085 (.37)	0.085 (.36)
0.15	0.144 (.49)	0.143 (.50)	0.087 (.33)	0.087 (.34)
0.20	0.157 (.49)	0.160 (.50)	0.088 (.32)	0.089 (.31)
	$L_b = -1/30, n = 30$			
−0.03		0.057 (.77)		0.057 (.77)
−0.01		0.059 (.56)		0.057 (.55)
0.00	0.062 (.53)	0.062 (.51)	0.058 (.50)	0.058 (.49)
0.05	0.088 (.51)	0.089 (.50)	0.059 (.38)	0.059 (.37)
0.10	0.111 (.50)	0.112 (.48)	0.060 (.32)	0.060 (.29)
0.15	0.132 (.49)	0.131 (.49)	0.061 (.26)	0.061 (.25)
0.20	0.147 (.47)	0.147 (.50)	0.062 (.23)	0.063 (.25)

Note. The τ is restricted to a lower bound of $-\sigma^2/n$. For data generated under the LME $\tau \geq 0$. LME is the linear mixed effects model, CSM is the covariance structure model, LM is the linear regression model.

Table 7.4: Estimates of τ, averaged across 1,000 replication, for different observations per cluster and $n = 10$ clusters under the LME.

	(1)	(2)
GEN	LME	CSM
ANA	LME	LME
τ	$\hat{\tau}$	$\hat{\tau}$
	$L_b = -1/10, n = 10$	
−0.09		0.095 (.80)
−0.07		0.097 (.68)
−0.05		0.098 (.61)
−0.03		0.100 (.57)
−0.01		0.105 (.53)
0.00	0.107 (.52)	0.107 (.53)
0.05	0.123 (.50)	0.124 (.52)
0.10	0.138 (.50)	0.139 (.50)
0.15	0.157 (.51)	0.156 (.50)
0.20	0.170 (.49)	0.168 (.51)
	$L_b = -1/15, n = 15$	
−0.06		0.079 (.76)
−0.04		0.080 (.62)
−0.02		0.083 (.57)
0.00	0.087 (.52)	0.087 (.52)
0.05	0.108 (.52)	0.106 (.51)
0.10	0.127 (.50)	0.126 (.49)
0.15	0.144 (.49)	0.143 (.50)
0.20	0.157 (.49)	0.160 (.50)
	$L_b = -1/30, n = 30$	
−0.03		0.057 (.77)
−0.01		0.059 (.56)
0.00	0.062 (.53)	0.062 (.51)
0.05	0.088 (.51)	0.089 (.50)
0.10	0.111 (.50)	0.112 (.48)
0.15	0.132 (.49)	0.131 (.49)
0.20	0.147 (.47)	0.147 (.50)

Note. LME is the Linear Mixed Effects model, and CSM is the Covariance Structure model.

tained more information about the intercept than captured by the LME. Specifically, when the correlation was zero, the LME slightly overestimated the p-value due to assuming that there was a small amount of correlation, while there was no correlation in the data. This was caused by the fact that the level of zero correlation is a lower bound under the LME. When the LM is used as model of analysis, the deflation and inflation of p-values is shown for increasing values of the correlation in each cluster. For $\tau > 0$, the underestimation of the p-values occurred when ignoring the correlation, which is shown under the LM. When the level of correlation is negative, there is a steep increase in the overestimation of the p-values for increasing negative values of the correlation under both the LME and LM. It can be concluded the LME performs as poorly as the LM, when it concerns a negative cluster correlation.

In Figure 7.2, the average SE estimates are shown for increasing values of the correlation, where the LM and the LME is used as model of analysis. For $\tau > .05$ correct SE estimates were retrieved under the LME, as it accounts for the positive correlation in the clusters. It can be seen that under the LM as model of analysis, the underestimation of the SEs became more severe for increasing values of the correlation. This can be seen from the increase in difference in SE estimates under the LM and the LME. For $\tau = 0$, the LM was the correct model of analysis, and in that case the SE estimate under the LME was slightly higher showing that the LME incorrectly assumed more correlation in the data leading to an overestimation of the SE. For $\tau < 0$, the SE estimates under the LME were more similar to those under the LM. However, even for small negative correlations, the SE estimates under the LME were higher than those obtained under the LM. The LME incorrectly assumed a positive correlation among clustered observations. Furthermore, for negative correlations the total variance reduced leading to a slight decrease in estimated SEs under both models.

Estimation of the correlation τ

The values of τ were estimated under the LME, and the data were generated under LME and under CSM. The τ values for data generated under LME were restricted to be greater than or equal to zero, where under CSM the τ values were also negative. The reported estimates of τ in Table 7.2 were averaged values across the 1,000 data sets. It can be seen that the estimated τ values under LME were biased, when the true τ was less than .05 or negative. In specific, the τ estimates were biased upwards, when the true value of τ was equal to zero. As can be seen in Table 7.2, the bias was smaller for larger values of the cluster size, because with an increasing sample size there was less sampling error. This

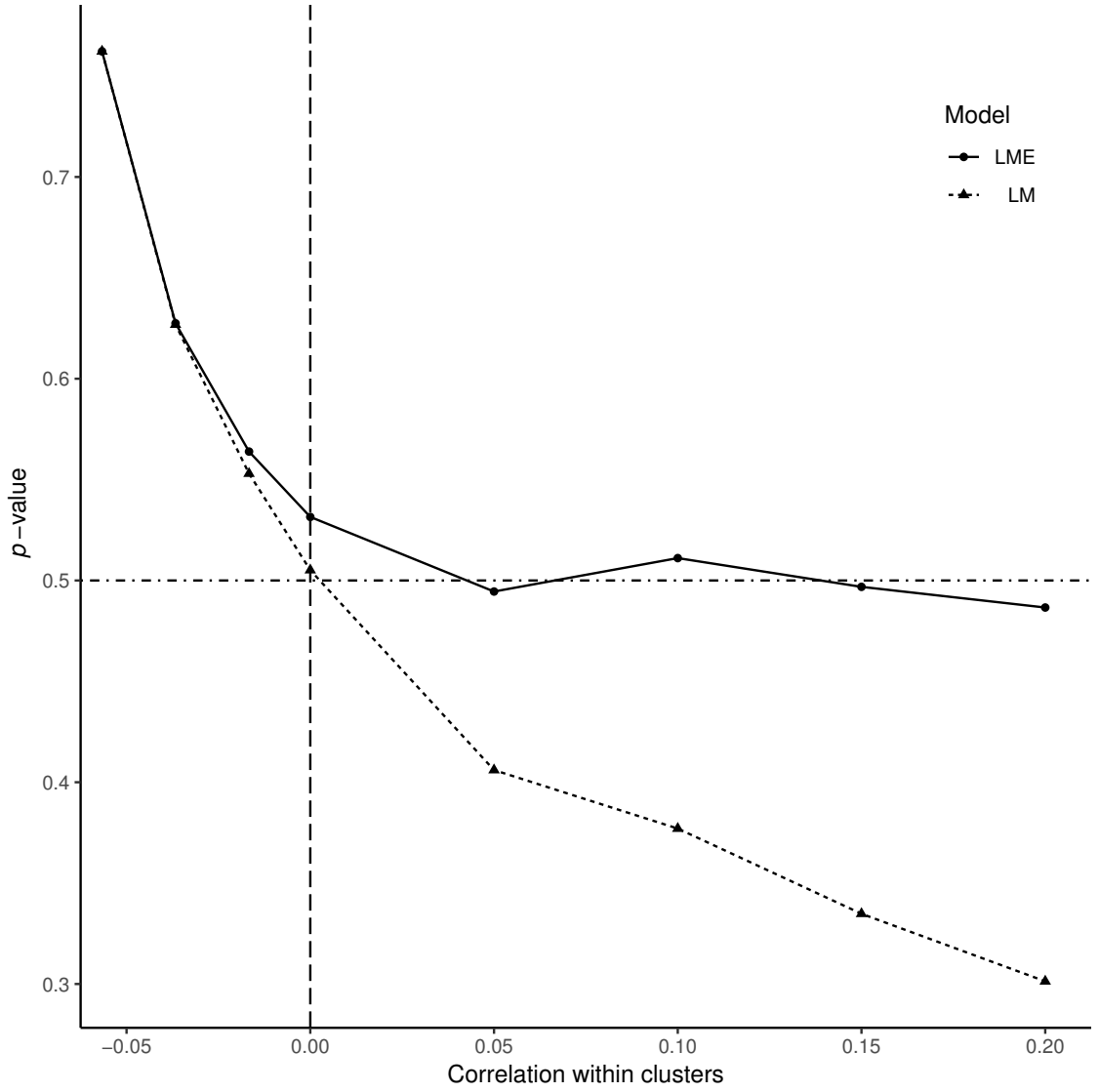


Figure 7.1: Estimated p-values for the intercept (averaged across 1,000 replicated data sets) for different levels of correlation among $n = 10$ clustered observations and $m = 10$ clusters for the LME and the LM.

led to less variability in the data, which means that more information about the fixed effect was included in the data. For a true positive correlation of $\tau = .05$ still some bias was found in the estimates for small cluster sizes. When the true positive τ values were greater than or equal to .10, the estimates of τ were not biased. This effect was equal for all cluster sizes. However, for small negative τ, small positive correlations were estimated under the LME, when the cluster size was small, and a zero correlation was estimated, when the cluster size increased to 30 observations.

Regarding the p-values, the ignorance of a negative ICC led to an overestimation of the p-values in an LME analysis for data generated under CSM. Furthermore, for an ICC greater than zero, the p-values were correctly estimated

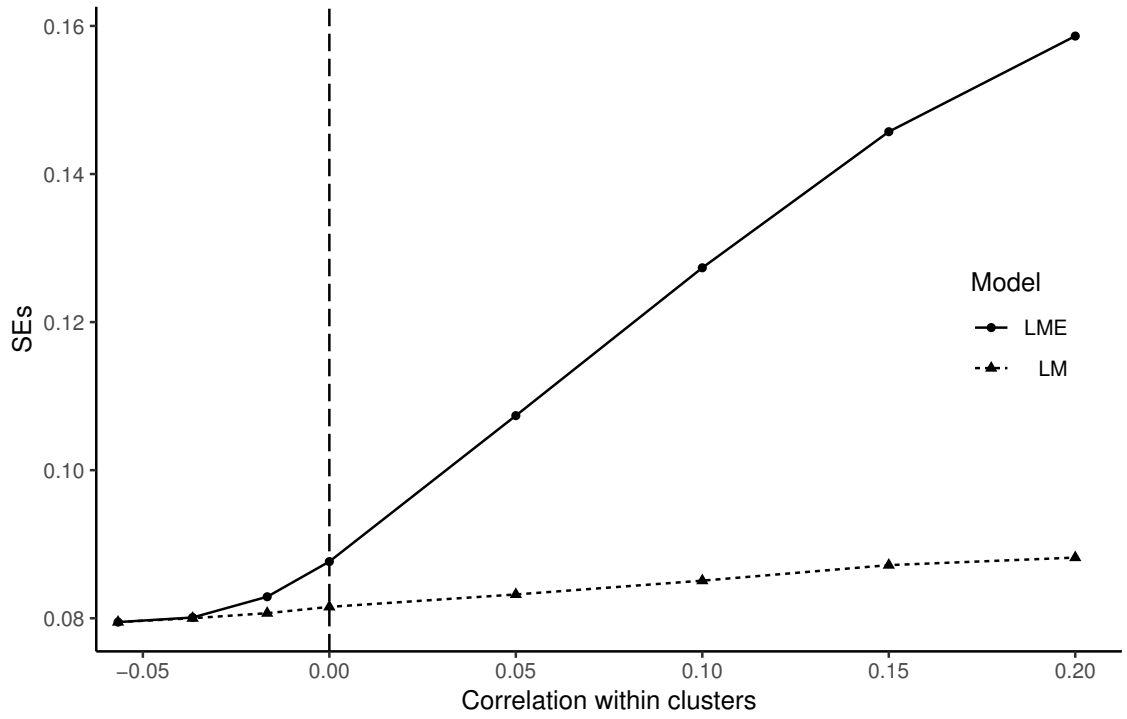


Figure 7.2: Estimated standard errors (SEs) for the intercept (averaged across 1,000 replicated data sets) for different levels of correlation among $n = 10$ clustered observations and $m = 10$ clusters under the LME and the LM.

at about .50 by the LME as model of analysis. Similarly, the relation between ICC and SEs correspond to the relation between τ and the SEs. The smaller the ICC, the smaller was the estimated SE.

Real data example

The data from Lamers et al. (2015) were used to study negative clustering effects. The dataset included 174 clients who were recruited through advertisements in Dutch newspapers and web-sites. Only participants who felt depressed and were expressed interest in writing about their life were included. Clients with no missing data were used leading to a total of 90 clients, denoted as $i = 1, \ldots, n$, who were equally divided over $m = 5$ counsellors, denoted as $j = 1, \ldots, m$ (leading to a balanced study design). The study had two treatment-arms: autobiographic writing condition (AW), and expressive writing condition (EW).

The AW condition was a life-review self-help intervention that consisted of homework assignments, divided over modules that had to be completed over the course of ten weeks. Clients communicated about their progress with trained counsellors through a weekly e-mail interaction. The EW intervention was based on the method of expressive writing by Pennebaker (1997). The method consisted of daily writing about emotional experiences, for $15 - 30$ minutes on $3 -$

4 consecutive days during one week. The available data included the pre- and post-therapeutic measurements, denoted as $t = 1, 2$, of the Center for Epidemiologic Studies Depression Scale (CES-D) score. The CES-D is a brief self-report questionnaire to measure severity of depressive symptoms in the general population. Higher CES-D scores indicated more depressive symptoms (20 items, range $0 - 60$, $\alpha = 0.78$).

Interest was focused on examining the influence of the counsellor on the health improvement of the clients. Therefore, a two-factor (LME) model was examined, with a (nested) factor client (clustering the pre- and post measurements), and clients were again clustered by counselors. The treatment indicator was included to examine differences in scores among the treatment groups. The LME represents this two-factor model, and is given by,

$$\begin{aligned} &Y_{ijt} = \beta_{0t} + \beta_1 \text{Treatment}_{ij} + \text{client}_{ij} + \text{counselor}_j + e_{ijt} \\ &e_{ijt} \sim N(0, \sigma^2) \\ &\text{client}_{ij} \sim N(0, \delta) \\ &\text{counselor}_j \sim N(0, \tau). \end{aligned} \tag{7.8}$$

The fit of this LME model led to estimation problems. The estimated covariance matrix was singular, and the variance component representing the clustering by counsellors was estimated to be zero (or less than zero, but `LME4` does not report information about negative variance component estimates; Smink, Fox, et al., 2019). It was concluded that the model with counsellor as a random effect was singular, and the model parameters could not be estimated.

The model can be rewritten as a multilevel SEM to allow for negative within-cluster correlations, by assuming that the pre- and post measurements of each client is multivariate normally distributed. The multilevel structure is defined by clients (factor variable) who are nested within counselors (factor variable). Then, the ML-SEM, referred to as M1, is represented by

$$\begin{aligned} &\mathbf{Y}_{ij} = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda}\boldsymbol{\eta}_{ij} + \mathbf{e}_{ij} \\ &\boldsymbol{\eta}_{ij} = \text{client}_{ij} + \text{counselor}_j \\ &\text{client}_{ij} \sim N(\beta_{1W} \text{Treatment}_{ij} I(Post), \delta) \\ &\text{counselor}_j \sim N(0, \tau). \\ &\mathbf{e}_{ij} \sim N(0, \sigma^2 \mathbf{I}_2), \end{aligned} \tag{7.9}$$

where the vector of loadings, $\boldsymbol{\Lambda}$, contains only ones, and the (within-counselor) treatment effect is defined for the post measurements.

Table 7.5: Real Data: Estimated ML-SEM parameters

Parameter			*Estimate*	*Std. Error*
		M1		
Fixed				
Intercept	Pre-		22.851	1.021
Intercept	Post-		18.129	1.021
Treatment	EW	(W)	−1.019	1.094
Treatment	EW	(B)	−2.436	1.864
Random				
Residual		(W)	21.611	3.222
Client		(W)	15.873	4.398
Counsellor		(B)	−1.110	0.343
		M2		
Fixed				
Intercept	Pre-		22.800	0.832
Intercept	Post-		17.625	0.832
Treatment	EW	(W)	20.500	0.715
Treatment	EW	(B)	15.675	0.715
Random				
Residual	AW	(W)	21.922	4.902
Residual	EW	(W)	17.422	3.896
Client	AW	(W)	14.463	6.553
Client	EW	(W)	9.865	4.894
Counsellor	AW	(B)	−1.088	1.525
Counsellor	EW	(B)	−0.852	1.083

Model M1 was fitted in Lavaan (version 0.6-5), and it reported convergence of the estimation algorithm with a warning that some latent variable variances estimates were negative. The fit indices did not indicate a misfit of the model; CFI=1 and RMSEA=0. In Table 7.5, the parameter estimates and standard errors are reported. It follows that on average the CES-D decreased from around 21.6 to 17.6 for those in the AW condition. Clients in the EW condition scored approximately 1.4 points lower on the post-measurement. However, the within-counselor treatment effect was not significantly different from zero.

The covariance among scores of clients of the same counselor was estimated to be around -1.159, which differed significantly from zero. The negative correlation made it impossible to investigate a homogenous treatment effect of counsellors. The negative correlation among observations clustered by the counsellor showed that scores differed substantially among clients who were treated by the

same counsellor. The counsellor differentiated clients in their treatment, where some clients benefitted much more from the treatment than other clients.

It was investigated of the negative correlation of clustered scores by counselors differed across treatment levels. Therefore, a multivariate distribution was assumed for the client scores in the EW and AW condition, where for each condition an intercept (pre and post-measurement), residual variance, client and counselor factor variance was defined. The parameter estimates of the multilevel SEM are given in Table 7.5 under the label model M2. It can be seen that on average the clients in the EW condition scored lower on the pre and post-measurement. When examining the intra-cluster correlation (ICC) of the factor variables, it follows that for the AW condition the ICC is around -8.1% (-1.088/(-1.088+14.463)), and for the EW condition around -9.5% (-.852/(-.852+9.865)). However, the differentiation in EW and AW counselor-specific cluster effects led to non-significant cluster (co)variance estimates. Note that the negative ICC shows that the counselor explained variability in client scores, but it led to more dissimilarity between clients. The treatment effect of a counselor varied across clients, where some clients benefitted from the treatment and others did not. A positive ICC would indicate a homogenous treatment effect, where all clients benefit from the treatment, but some more than others. A more negative ICC leads to a greater distinction between clients treated by the same counselor who benefit and who do not benefit from the treatment. Clients in the EW condition were more likely to differentiate in the effect of their treatment than those in the AW condition but the differentiation in clustering between EW and AW clients was not significant.

The main conclusion was that counselors differed in their treatment of the clients, leading to dissimilarity between clients treated by the same counselor. This negative clustering effect indicated that the individualized treatment by counselors only worked for some clients and not for other clients. The differentiation in treatment effects between clients (most likely) led to a non-significant main difference between the AW and EW condition.

Discussion

The purpose of this study was to confirm that estimates of the ICC were biased upwards for true small positive and negative values. The results confirm that the ignorance of (small) positive correlation within groups leads to an increase of Type-I error, p-values that are biased downwards and CIs that are too narrow. The results for ignoring true positive correlation within clusters vali-

date earlier research. The inflation of Type-I errors of this study correspond to those reported by Barcikowski (1981). Regarding the different conditions, when ignoring a positive correlation within groups, the results of this study also confirm earlier research. As reported by Barcikowski (1981) and Dorman (2008), the smallest bias in Type-I error was found for a smaller number of observations per group. When the number of observations per group increased, the inflation of the Type-I error also increased. Furthermore, when increasing the positive correlation, the Type-I error was increasing as well. Furthermore, the results show that, in general, only the intercept parameter and not the slope parameter was affected by ignoring a common correlation within clusters. Effects of slope variables are not affected, when ignoring a common correlation the distortion only concerns the intercept parameters.

Regarding the estimates of the correlation, the results show the negative effects of restricting the correlation among clustered observations to be positive. Already for small positive correlation ($\tau < .05$) the correlation was overestimated. The bias was larger when the correlation was very small positive or negative. However, for a positive correlation within groups larger than .05, the LME model gave correct estimates, which validates the accuracy of the LME for clustered data with positively correlated observations.

Next to ignoring a positive correlation, the ignorance of a (small) negative correlation within groups was considered. The results showed opposite effects compared to ignoring positive correlation within groups: a deflation of Type-I errors, p-values that are biased upwards and overestimated SEs (i.e. CIs that are too wide). The deflation of the Type-I error, when ignoring a negative correlation has been mentioned by other researchers (Barcikowski, 1981; Rosner & Grove, 1999), but without quantifying the negative effects of ignoring the common negative correlation. Current findings indicate that when clustered observations are negatively correlated, the data is more informative than it would be under independent sampling. This study adds that smaller biases of the Type-I error and the SEs occur, when increasing the number of observations in a cluster. The increase in sample size leads to smaller sampling error and more accurate estimates. In addition, the bias of the Type-I error is examined, where a deflation of the Type-I error by 2% can already occur for a true negative ICC of $-.01$. Furthermore, the results indicate that p-values were biased upwards when negative correlation within groups was ignored. For a larger number of observations per cluster, even the ignorance of a very small negative correlation within groups can lead to substantial bias of the p-value.

With respect to the results of the current study, it is recommended to be aware of the fact that a negative correlation within groups can occur. The advise is to

be very cautious with clustered data that possibly contains a common negative correlation among observations within groups. Researchers need to be aware of the fact that the maximum likelihood estimate of the between-cluster variance, representing the covariance among clustered observations (see Equation 7.2) can become negative. Common software packages for mixed effects models, such as the `lme4`-package, restrict the random-effect variance estimate to be positive. This means that the correlation among clustered observations is restricted to be positive even when the true correlation is negative. This study showed that this can lead to a deflation of Type-I errors, p-values that are biased upwards and SEs that are overestimated. In addition, a large number of observations per cluster does not compensate for bias. It is important to notice that with 30 observations per cluster a very small true negative correlation of $-.008$ can lead to an inflation of the p-value with 5%. It is also noted that for a smaller number of observations per cluster, the correlation can become more negative than with a larger number of observations. Thus, the bias of the p-value is smaller when the true negative correlation between observations is small. However, in that case the correlation can become more negative which can lead to more bias.

The findings of this study are limited in their application because, until now, there is no proper tool known for detecting negative correlation within groups. In case researchers are not aware that there might be negative correlation, the LME model automatically assumes the correlation to be positive and close to zero. This leads to bias that may not be recognised. In future research, it is essential to find a tool or strategy to detect and model negative correlation within groups. Finally, if the data have negative correlations the LME model will assume that the observations are independently distributed, which makes the LME model an inappropriate tool for analysing clustered data with negative correlations. Note that bias in the estimated precision of the slope parameter can occur when ignoring correlation among clustered observations. Consider a random intercept-slope model,

$$\mathbf{y}_j = \beta_0 + \beta_1 \mathbf{X}_j + u_{0j} + u_{1j}\mathbf{X}_j + \mathbf{e}_j, \tag{7.10}$$

where both random effects are normally distributed, $u_{0j} \sim \mathcal{N}(0, \tau_0)$ and $u_{1j} \sim \mathcal{N}(0, \tau_1)$. Subsequently, the implied covariance matrix for cluster j is represented by $\mathbf{\Sigma}_j = \sigma^2\mathbf{I}_n + \tau_0\mathbf{J}_n + \tau_1\mathbf{X}_j\mathbf{X}_j^t$. The estimated precision of the intercept contained bias, when ignoring the common correlation τ_0 in cluster j. In the same way, the estimated precision of the slope parameter will contain bias, when ignoring the common correlation τ_1 modified by the outer product of $\mathbf{X}_j$. However, more

research is needed to quantify bias in the precision of the common regression effect, when ignoring cluster correlation implied by a cluster-specific regression effect.

What can be learned from our study?

We conclude with a discussion of what we found to be the most important lessons (that can be learned based on our study), and we give an updated overview of well-known multilevel modelling (golden) rules by including the findings concerning negative clustering effects. This reflection highlights the importance of understanding negative clustering effects.

Lesson 1: Negative correlations among clustered observations can –and do– occur in practice

Although this has been addressed in the introduction, it is important to stress that –although simulated data were used in the current study– negative correlations can and do occur in practice. Several practical examples were discussed, when the correlation τ is negative. Pryseley et al. (2011) and Kenny et al. (2002) mention several examples: when individuals compare for a scarce (and fixed) set of resources, the speaking time of one individual is at the expense of another individual (i.e. '*one's pain is the other's gain*'). Litter mates are also negatively correlated in terms of food, water and living space. In addition to that, it has been shown that *Bayesian Covariance Structure Modelling* (BCSM; Klotzke & Fox, 2019a, 2019b) can be used to simulate these negative correlations in clusters. In doing so, a simulation study was developed where observations are negatively clustered and data were simulated that others, most notably Kenny et al. (2002) and Oliveira et al. (2017), and Pryseley et al. (2011) described.

Lesson 2: The ICC has multiple interpretations

It is important to be aware of the multiple interpretations of the ICC, because viewing it as the proportion of explained variance due to clustering restricts the ICC to only positive values (Oliveira et al., 2017). It is good to stress that the ICC remains a measure of correlation, and correlations can become negative. A negative τ leads to a negative ICC, thus both values can become negative. Researchers need to be aware that negative correlation between observations in clustered data may occur, and that negatively correlated clustered data are not (always) a modelling error.

Lesson 3: The LME ignores negative associations in clusters

The LME (e.g., multilevel model, random effects model) is an inappropriate tool for analysing clustered data with negative correlations between observations within clusters. The LME is a well-established tool to model the dependence structure of the clustered data using random effects. However, this is only possible for positively correlated observations, but –as shown– the random effects cannot model negative correlations, and consequently will ignore the negative dependence between observations. When the LME is used for analysing data with negative correlations within clusters, the model assumes the observations to be independently distributed. Therefore, the LME is an inappropriate tool for analysing clustered data with negative correlations within clusters.

Lesson 4: Do not fix the ICC or τ to 0

Fixing a negative ICC to zero forces the statistical model to ignore the clustered structure in the data. Then, the observations are assumed to be independently distributed while they are negatively correlated. It is shown that if a common negative correlation in clusters is ignored, the Type-I errors are deflated, the SEs are too large, and the p-values are also overestimated. Fixing the ICC to 0 forces the statistical model to assume that data are independently distributed. Researchers should always account for a non-zero ICC, positive or negative. Huang (2018) stated that "literally too many to list" suggested that it is not necessary to account for relatively small (positive) ICC values. Given the fact that ignoring a negative correlation can also have a large impact, it is stressed that small positive or negative ICCs cannot be ignored. For instance, when having ten observations in each group, ignoring an ICC of only $-.029$ ($\tau = -.03$) leads to an deflation of the Type-I error of around 6%. The occurrence of a Type-I errors quickly increases: an ICC of $-.047$ ($\tau = -.05$) deflates the Type-I error with 11%.

Lesson 5: An increase in cluster size increases the bias

When the number of observations in each cluster n tends to get larger, the bias of the Type-I error and the SEs increases. The results indicate that p-values were biased upwards (downwards) when negative (positive) correlation within groups was ignored, and the bias increased with an increasing cluster size. When assuming more clustered observations to be independently distributed, while they are correlated, the bias will increase since an incorrect assumption is made about a larger data set.

Lesson 6: Negatively correlated clustered observations cannot be ignored

It was shown that ignoring a small negative ICC leads to serious bias. It can also be argued that a cluster sample with negatively correlated observations contains more information than a simple random sample of the same size. The variance inflation factor (VIF; i.e. design effect; Kish, 1965) can be used to explain the impact of the ICC (ρ). In the random selection of equal clusters, the VIF can be expressed as a function of ρ;

$$VIF = 1 + (n - 1)\rho, \tag{7.11}$$

where n is the number of observations in a cluster. The VIF represents the ratio of the actual sample variance to the variance of a simple random sample (i.e. independently distributed observations). In cluster sampling, when the ICC is positive, the VIF is greater than one. There is a loss in precision (i.e., increase in sample variance) due to the homogeneity of observations within each cluster. A simple random sample of the same size contains more information than a cluster sample with positively correlated observations. However, when the observations are negatively correlated, leading to a negative ICC, the VIF will be smaller than one, but greater than zero given that the ICC is greater than $-1/(n-1)$. In that case, the cluster sample leads to a gain in precision in comparison to simple random sampling. The cluster sample, with negatively correlated data provide more information than a simple random sample of equal size.

Conclusion

Just as small (positive) clustering effects mandate adjustment, negative clustering effects also require the attention of researchers. It is shown that ignoring negative clustering leads to deflated Type-I errors, an overestimation of the SEs, and overestimated p-values. The LME is an inappropriate tool for the analysis of negative clustering in data and can only handle positive clustering. Although it seems obvious, the LME should not be used when there is no clustering.

8 Assessing an Alternative for 'Negative Variance Components': A Gentle Introduction to Bayesian Covariance Structure Modelling

Abstract

The multilevel model (MLM) is the popular approach to describe dependences of hierarchically clustered observations. A main feature is the capability to estimate (cluster-specific) random effect parameters, while their distribution describes the variation across clusters. However, the MLM can only model positive associations among clustered observations, and it is not suitable for small sample sizes. The limitation of the MLM becomes apparent when estimation methods produce negative estimates for random effect variances, which can be seen as an indication that observations are negatively correlated. A gentle introduction to Bayesian Covariance Structure Modelling (BCSM) is given, which makes it possible to model also negatively correlated observations. The BCSM does not model dependences through random (cluster-specific) effects, but through a covariance matrix. We show that this makes the BCSM particularly useful for small data samples. We draw specific attention to detect effects of a personalized intervention. The effect of a personalized treatment can differ across individuals, and this can lead to negative associations among measurements of individuals who are treated by the same therapist. It is shown that the BCSM enables the modeling of negative associations among clustered measurements and aids in the interpretation of negative clustering effects. Through a simulation study and by analysis of a real data example, we discuss the suitability of the BCSM for small data sets and for exploring effects of individualized treatments, specifically when (standard) MLM software produces negative or zero variance estimates.

Keywords: Bayesian Covariance Structure Modelling (BCSM), individualized treatment, negative variance estimates, negative clustering effects, multilevel modelling

This chapter has been submitted as: Fox, J. P., & Smink, W. A. C. (2020).

Introduction

Data are so often plagued by observations that are correlated (i.e. clustered, not independently sampled) that it is difficult to overstate the importance of multilevel models. This family of statistical models aids researchers in understanding the clustered –or hierarchical– structures in their data (i.e. 'groups within groups', 'non-independent data', or 'hierarchical data'). In the multilevel modelling framework, dependences among observations are expressed as a covariance, which are modelled through a random effect, also known as a latent variable. The variance of the random effect determines the strength of the correlation among clustered observations. For a small variance, clusters are similar to each other and observations within a cluster do not correlate highly. With high random effect variance, the cluster-specific parameters show large differences and observations within each cluster are much more alike than those from different clusters.

However, modelling the clustering effect (i.e. magnitude of the –positive– correlation between observations) as the variance of a random effect also introduces a great –and relatively unknown– shortcoming of multilevel models. Multilevel models impose the restriction that within-cluster correlations should be positive, since the variance of a random effect is restricted to be positive. However, correlations are not restricted to positive values only, they can also be negative or zero. Indeed, although not widely known, negative correlations among clustered observations can also occur (Kenny et al., 2002), but the multilevel model cannot assess these effects.

The multilevel model describes within-cluster *similarity* through a *positive* random effect variance. From this perspective to describe within-cluster *dissimilarity* with a multilevel model, a negative random effect variance would be required. This has led to interest in estimating and interpreting negative variance components and identifying negative clustering effects (El Leithy et al., 2016; Kenny et al., 2002; Molenberghs & Verbeke, 2007, 2011; Nelder, 1954; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003). These effects remain unknown to the sheer majority of scientific community. Furthermore, methods to apply multilevel modeling techniques for analysing within-cluster dissimilarities are limited, and have not been expanded to address more complex clustered data. Molenberghs and Verbeke (2011) discussed a marginal model representation of the random intercept model (by integrating out the random effect), and Snijders and Kenny (1999) adjusted a multilevel model with correlated dummy variables to describe negative within-cluster dependence.

Recently, Fox et al. (2017) and Klotzke and Fox (2019a), and Klotzke and

Fox (2019b), developed a rigorous new Bayesian modeling approach for clustered data, referred to as *Bayesian Covariance Structure Modelling* (BCSM). Based on multilevel modeling principles, BCSM can describe similarities and dissimilarities. We intend to give a gentle introduction to the BCSM here, and stress possibilities of the framework to deal with negatively and positively correlated observations. BCSM is a relatively simple and flexible covariance structure modelling approach, which avoids several restrictions of the popular multilevel models.

In short, in the BCSM approach, a dependence structure is not *indirectly* modelled through random effect parameters. The dependence structure is *directly* modelled by specifying a structured covariance matrix. This structured covariance matrix represents the correlations among clustered observations to account for the fact that the observations are not independently distributed. Both modelling approaches are discussed by considering the one-way random effects model (i.e. random intercept model). This relatively simple model is used as a vehicle to introduce the BCSM and its potential for modelling clustered data. Subsequently, BCSMs for more complex dependence structures, for instance a two-way (nested) structure, are described. The potential of the BCSM is supported by a straightforward Gibbs sampling method to estimate all model parameters, where (co)variance parameters can be directly sampled from inverse-gamma distributions.

We organized the remainder of the this paper as follows. We give a gentle introduction to modelling clustered data using random effects. Then, we introduce BCSM, emphasizing –not technical rigour, but– understanding of the framework. We aim to convince those who are potentially interested in BCSM about the advantages of the approach by reporting on the results of our extensive simulation study, which shows that BCSM can –indeed– detect positive as well as negative within-cluster dependence. In addition to that, we show that (very) small variance components of random effects (i.e. that are very close to zero) can be accurately estimated. We also demonstrate that BCSM can describe efficiently complex dependence structures with a few (co)variance parameters making it particularly useful for small data samples. We assess a real-data example, where differences in pre- and post-intervention depression scores between two treatment arms are examined, while accounting for a clustering by counsellors. The example illustrates why a negative clustering effect cannot be ignored, as these effects also occur in practice. Our overall goal is to not only discuss the statistical importance of negative clustering effects, but to show how negative effects should be interpreted. We will argue why clinical practitioners and psychotherapy researchers (and all others who are interested in knowing *what*

works when for whom; Norcross & Wampold, 2011; Smink, Sools, van der Zwaan, et al., 2019; Tasca et al., 2015)[1] are – in fact– interested in interpreting negative clustering effects. Finally, the specific features of BCSM are discussed, including its strengths and limitations.

Modelling Clustered Data

It is clear that various complex forms of clustering and hierarchical organisations arise naturally in a multitude of settings in psychological research. What all these settings have in common, is that –in their fundamental form– each multilevel model consists out of a *within* and a *between* cluster component. With two levels, the multilevel model (MLM) defines separate probability distributions for the clusters and for individuals within these clusters. Under the cluster-sampling design, clusters are assumed to be independently sampled from a population, and individuals are assumed to be independently sampled from each cluster. This two-stage sampling design is represented in the MLM, which specifies a probability distribution for the cluster-specific parameters (i.e. random effects) and a probability distribution for the lower-level observations. Following the properties of the two-stage sampling design, observations from individuals are assumed to be conditionally independently distributed given the cluster-specific (random effect) parameters.

Cluster-specific parameters (i.e. random effects) are often used to model clustered data and they are included in the mean regression component. When conditioning on the cluster-specific parameters, the observations within the cluster can be assumed to be independently distributed. In doing so, the assumption of independence is no longer violated, as the correlation of the clustered data is bypassed through inclusion of these cluster-specific parameters. This technique is used in the very popular models as hierarchical linear regression models (Raudenbush & Bryk, 2002b), random effect models (Longford, 1995), multilevel models (Goldstein, 2011; Snijders & Bosker, 2012), and linear mixed effect models (McCulloch et al., 2008; Verbeke & Molenberghs, 2009). We use the term MLM to represent these type of (conditional) models. It is (also) good to note that this class of models can be extended further: in the latent class models (Vermunt, 2008), or mixture models (McLachlan & Peel, 2000), observations within each latent cluster are also assumed to be conditionally independently distributed given the cluster-specific parameters.

We use the one-way random effects model (i.e. random intercept model) to

[1] Smink, Sools, van der Zwaan, et al. (2019) is chapter 2 of this thesis.

demonstrate the modelling of the within-cluster correlation with a (random effect) variance parameter. Through standard equations, we will show now that the within-cluster correlation is restricted to be positive, since the random effect variance cannot be negative. The one-way random effects model is most commonly used for describing continuous data that are clustered in one way. Without making an explicit distinction between a random variable and a realized value, the outcome y_{ij} is the j-th observation in the i-th cluster and expressed as the sum of the general mean, μ, random effect α_i, and residual error e_{ij},

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + e_{ij}, \\ \alpha_i &\sim N(0, \tau), \\ e_{ij} &\sim N(0, \sigma^2). \end{aligned} \tag{8.1}$$

The within-cluster error variance, σ^2, represents the variation in observations in each cluster i given the random effect α_i. Aside from the term random effect, the α_i is also referred to as the *blocking* factor, *grouping* factor, or the *treatment* factor. The random effect is assumed to be normally distributed with mean zero and variance τ. The random effect variance τ represents the variation in random intercepts across clusters, and is often referred to as the between-cluster variance. Indeed, this restricts the τ to positive values only. The BCSM approach will relax this restriction by introducing a different representation of the model, and will later show why this is relevant.

The random effect variance parameter is not presented in a squared notation, since this variance parameter also represents the covariance among clustered observations. A covariance parameter is not restricted to be positive, but squared terms always are. The relation between the covariance and τ becomes immediately apparent when considering the covariance between the two clustered observations j and l in group i, which is represented by

$$\begin{aligned} Cov\left(y_{ij}, y_{il}\right) &= Cov\left(E\left(y_{ij} \mid \alpha_i\right), E\left(y_{il} \mid \alpha_i\right)\right) + E\left(Cov\left(y_{ij}, y_{il} \mid \alpha_i\right)\right), \\ &= Cov\left(\mu + \alpha_i, \mu + \alpha_i\right) + 0, \\ &= Cov\left(\alpha_i, \alpha_i\right) = Var\left(\alpha_i\right) = \tau. \end{aligned} \tag{8.2}$$

Then, the variance of an observation equals

$$\begin{aligned} Var\left(y_{ij}\right) &= Var\left(E\left(y_{ij} \mid e_{ij}\right)\right) + E\left(Var\left(y_{ij} \mid \alpha_i\right)\right), \\ &= \sigma^2 + \tau. \end{aligned}$$

The intra-class correlation (ICC) represents the proportion of variance explained

by the clustering. This is represented by

$$\rho = \frac{Cov(y_{ij}, y_{il})}{Var(y_{ij})} = \frac{\tau}{\sigma^2 + \tau}. \tag{8.3}$$

Under standard MLMs, the ICC is restricted to be positive since both the numerator (τ) and the denominator ($\sigma^2 + \tau$) are variance parameters. Generally, the interpretation of ρ stops here, as the general tendency is to think that an ICC cannot be negative, restricting ρ to lie between zero and one (see for example Eldridge et al., 2009; Huang, 2018).

However, the covariance component in the numerator in Equation (8.3) could also be negative if τ represents the covariance among clustered observations, and not also the random effect variance. It is this double function of the random effect variance parameter τ that restricts the covariance among clustered observations, and the ICC, to be positive.

Examples of Negative Clustering

We discuss several examples where researchers either encountered negative ICC values, or where they could be expected. Note that it is currently difficult to give a literature overview: researchers do not report on negative ICCs, nor that it is well-known that these values in fact occur, and the common statistical software packages do not allow for negative associations between clustered observations. We visualize depict the following examples in Figure 8.1.

Multidisciplinarity

Nowadays in science –but also in society and everywhere where people collaborate– it is increasingly important to think, act and create across boundaries. Therefore, the cultural, ethical or scientific background of one individual should differ from that of the others, which stimulates the dissimilarity in a group. This is a pattern that can also be seen in the police force, in politics, and education. Other forms of diversity in a group are co-morbidity, the number of co-morbid diseases increases with age, or through smoking, development of dementia and other diseases, human migration.

The boomerang effect

Another factor that can create negative clustering dependencies is referred to as the 'boomerang effect' (Kenny et al., 2002): one set of observations may influence the other observations in the cluster to be different. Figure 8.1 gives an

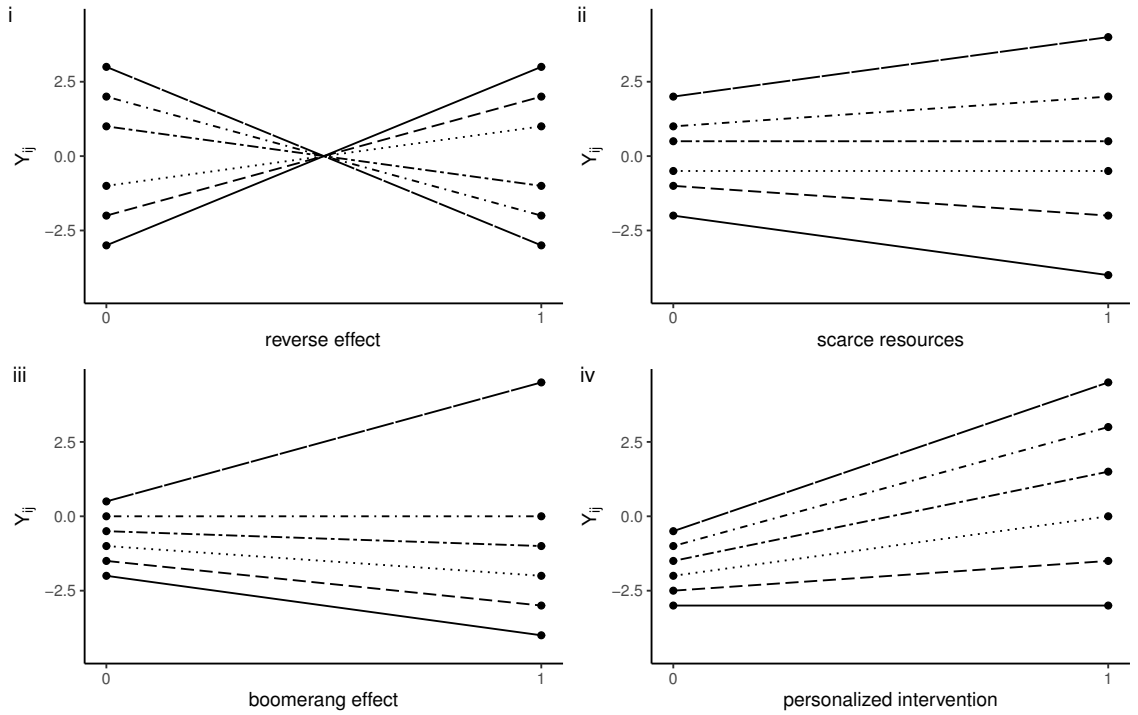


Figure 8.1: Four examples where the cluster variance τ is negative (scenario i - iv). Outcome Y_{ij}, observed two times (e.g., repeated measures, 0 for the pre-, and 1 for the post-observation). Scenario i shows a reverse effect, all pre-observations are reversed ('reverse effect') at the post-measurement; ii shows a setting where individuals are competing for scarce resources (e.g., one's pain is the other's gain); iii shows how other group members decrease as a reaction to the increase of another group member ('boomerang effect'; Kenny et al., 2002); and iv shows that clustered individuals have dissimilar trajectories related to the personalized intervention.

illustration of such pattern: in a pre-test post-test (or repeated measurements) design, rather than individuals behaving more similar, one (or a small set of observations) increases dissimilarity in a cluster. For example, in group or family counselling, the behaviour of a narcissistic individual might decrease the self-worth of others. A similar pattern might be noticed with a pessimistic individual: one pessimistic individual may increase the mood of others. In Figure 8.1, this boomerang effect is shown for a repeated measures setting with two measurement occasions, where the dissimilarity is increasing in the cluster.

viii

Competing

Another source that can cause negative clustering effects is competing, which was suggested by Pryseley et al. (2011). Figure 8.1, an illustration is given of how competing can increase the variances among the mebers of a group. Individuals often compete for the allocation of scare resources within the same group. The

examples suggested by Pryseley et al. (2011) are litter mates, division of a fixed reward, speaking time, and leadership. In Figure 8.1, this phenomena is referred to as *one's pain is another's gain*.

Personalized interventions

The typical situation in studies of psychotherapy process and outcome is that one counsellor treats several clients (Baldwin & Fellingham, 2013). When the clients who see the same counsellor are more similar to each other than those clients who are treated by different counsellors, outcomes of clients with the same counsellor are expected to be positively correlated. The counsellor treats clients in a similar way, which leads to a common positive correlation among the treated clients. Although the efficacy – or clustering effect – of the counsellor is well-known to be important, it is not always assessed. Doing so is straightforward in the MLM approach (Baldwin & Fellingham, 2013; Kenny & Hoyt, 2009; Marcus et al., 2009; Raudenbush, 2001). However, when the counsellor provides a personalized treatment, the effects of each treatment can differ substantially across clients. Personalized interventions are designed for the individual (Smink, Fox, et al., 2019):[2] *what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of *circumstances* (Paul, 1967, p. 111)? As a result, dissimilarity in a counsellor's client group can occur when for some individuals the personalized treatment works well but not for others. This can lead to a negative correlation among the treated clients of a counsellor. In fact, a negative correlation would indicate that some clients benefit highly from the personalized treatment, where for others positive treatment effects are more difficult to realize. The negative correlations also provide information about the counsellor who is able to improve the treatment of clients through personalization leading to dissimilar client results, since clients still respond in different ways to a personalized treatment.

The Bayesian Covariance Structure Model

The general idea of BCSM is to model directly the dependence structure of the data, and not indirectly through random effect parameters. This dependence structure can be implied by random effects. The BCSM is a more general approach for clustered data, since it can also identify a negative dependence structure and a dependence structure implied by non-identifiable random effects. BCSMs have been developed for different applications to deal with complex cor-

[2] Smink, Fox, et al. (2019) is chapter 6 of this thesis.

related data structures (Fox et al., 2017; Klotzke & Fox, 2019a, 2019b; Mulder & Fox, 2019).

Consider the error terms α_i and $\mathbf{e}_i = (e_{i1}, \ldots, e_{in})$ to describe the dependence structure for the clustered observations. The error component for cluster i, $\mathbf{E}_i = \alpha_i + \mathbf{e}_i$, is assumed to be multivariate normally distributed, where the covariance matrix comprehends the common covariance among the clustered observations (Equation 8.2) on the non-diagonal and the total variance (Equation 8.3) on the diagonal. It follows that,

$$\begin{aligned} \mathbf{y}_i &= \mu + \mathbf{E}_i, \\ \mathbf{E}_i &\sim N(0, \boldsymbol{\Sigma}), \end{aligned} \tag{8.4}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \tau & \tau & \ldots & \tau \\ \tau & \sigma^2 + \tau & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau & \ldots & \tau & \sigma^2 + \tau \end{bmatrix}. \tag{8.5}$$

Under the BCSM, parameter τ is no longer a variance parameter and only represents the common covariance among clustered observations. The τ is a covariance parameter and not a variance parameter. This has three important implications: 1) τ can now also be negative, 2) zero is no longer the boundary value for τ, and 3) τ is not estimated as the random intercept variance. Indeed, negative values for τ are now perfectly acceptable, since this merely corresponds to the occurrence of *negative* within-cluster correlation. The implications of negative clustering effects will be discussed later. The only requirement is that the covariance matrix is positive definite, which is the case when $\tau > -\sigma^2/n$ (which will be shown later).

Type of Dependence

It is straightforward to represent the covariance matrix in matrix notation. Assume that each cluster i has n observations, then $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n + \tau\mathbf{J}_n$, where the $\mathbf{J}_n$ is a matrix of dimension n with all elements equal to one and $\mathbf{I}_n$ is the identity matrix of dimension n. The dependence structure of this covariance matrix $\boldsymbol{\Sigma}$ is straightforward: if there is no clustering in the data, the covariance τ is not present (e.g. $\tau = 0$). If τ is positive, the observations are assumed to be positively correlated and the dependence structure is similar to that of the random intercept model in Equation (8.1). If τ is negative, the observations are negatively correlated within a cluster, a dependence structure that cannot be represented

by a random intercept model. Thus, the BCSM elegantly represents three nested models depending only on the sign and value of the covariance parameter. Indeed, the BCSM simply extends the range of possible values to include zero and negative values, without changing the interpretation of positive values.

Multiple Types of Dependence

In our real data example, for each client a pre-intervention and post-intervention score was observed, and clients were treated by counselors. Thus, observations were clustered by clients (type A clustering), who were again clustered by counselors (type B clustering). The BCSM can be extended to describe any additional type of clustering. To illustrate this, we consider our real-data design, where observations were clustered according to type A, as described in Equation (8.1), and that those clustered observations are again clustered according to type B. In the two-way random effects model, a random effect $\beta_{(i)j}$ can be introduced that represents the clustering of observations according to type B, which is represented by

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_{(i)j} + e_{ijk}, \\ \alpha_i &\sim N(0, \tau_a), \\ \beta_{(i)j} &\sim N(0, \tau_b), \\ e_{ijk} &\sim N(0, \sigma^2). \end{aligned} \tag{8.6}$$

It follows that objects in cluster i are nested (type A), where the τ_a represents the common dependence among the clustered observations. Within each cluster i, observations in each cluster ij are again nested (type B), where the τ_b represents the dependence among those clustered observations. The random effect variance parameters τ_a and the τ_b represent the dependence among clustered observations but are both restricted to be positive.

In the BCSM for this two-way (nested) structure, the dependence structure is directly modelled. To be complete, the covariance matrix is given for this design. Let b clusters of type B each of size n be nested within the cluster of type A, with in total a type-A clusters. Then, the BCSM covariance matrix is represented by

$$\boldsymbol{\Sigma} = \left(\mathbf{I}_{nb}\sigma^2 + \mathbf{J}_{nb}\tau_a\right) + \left(\mathbf{I}_b \otimes \mathbf{J}_n\right)\tau_b. \tag{8.7}$$

The covariance matrix of the one-way clustering is extended with an extra component that displays the nesting of observations in type B clusters. The Kronecker product $\otimes$ is needed to define which of the observations in each cluster

A are again nested according to cluster B. It states that the b blocks of n observations are clustered with a common dependence of τ_b. The BCSM for the two-way clustered data is represented by Equation (8.4) with the covariance matrix defined in Equation (8.7).

In the BCSM any type of clustering is directly modelled through the structured covariance matrix, and this covariance matrix can represent multi-way structured data. Furthermore, a hybrid version is also possible, where the mean component also includes random effect parameters. For instance, a hybrid version of a two-way BCSM can be defined by including the random effect $\beta_{(i)j}$ in the mean term with the structured covariance matrix of the one-way model in Equation (8.5). The BCSM represented in Equation (8.4) is also easily extended to include explanatory variables with fixed effects. Let $\mu = \mathbf{X}_i\boldsymbol{\beta}_f$, the (design) matrix $\mathbf{X}_i$ contains the explanatory variables for cluster i and the $\boldsymbol{\beta}_f$ represents the regression effects of the variables.

Advantages of BCSM over MLM

To summarize the previous section: the BCSM is a novel Bayesian modelling framework in which the covariance structure of a (complex) dependence structure is directly modelled. This makes the BCSM more flexible and more general than standard MLMs. We give an overview of the specific features of BCSM in comparison to MLM. Next to a theoretical discussion of the advantages of BCSM, we specifically designed our simulation and real-data study to provide more evidence in support of these claims.

Modelling Negative Clustering Effects

Negative correlations among clustered observations cannot be modelled with the MLM. In the MLM, a positive correlation is modelled through a shared group-specific effect among the group members. This modelling concept cannot be translated to model negative dependences, since sharing a common component always leads to a positive association. The BCSM has been developed with the purpose to model in a similar way positive as well as negative correlations among clustered observations, while using a common dependence structure across groups.

Although a well-known and widely applicable statistical model for negatively correlated clustered data is lacking, the negative effects of ignoring negatively correlated clustered data has been mentioned in the literature. Ignoring a positive correlation in the data leads to an increase of the Type-I error, p-values

that are biased downwards and confidence intervals that are too narrow. Standard errors of fixed regression effects are smaller than they should be, feigning a precision of the estimates that is not actually supported in practice, leading to spurious and erroneous results of statistical significance (Kenny et al., 1998). For relatively small clustering effects (i.e. for small values of the ICC), Barcikowski (1981) showed that for instance an ICC of 0.05 and 100 observations per group already inflates the probability of a Type-I error to 0.43. Next to ignoring a positive correlation, the ignorance of a (small) negative correlation within groups leads to opposite effects compared to ignoring positive correlation within groups: a deflation of Type-I errors, p-values that are biased upwards and overestimated SEs (i.e. confidence intervals that are too wide). This deflation of the Type-I error, when ignoring a negative correlation has been mentioned by other researchers (Barcikowski, 1981; Rosner & Grove, 1999). Nielsen et al. (2020)[3] also quantified in detail the negative effects of ignoring the negative correlation.

There is an apparent risk of ignoring dissimilarity (i.e. negative correlations) within clusters. Even the smallest dissimilarity between clustered observations can seriously inflate the probability of a Type-I error. Kenny et al. (2002) argue along the same lines: if *positive* clustering effects can cause various statistical problems, then so do *negative* clustering effects. It is well known that when the ICC is greater than zero, which often occur in psychology (Hox et al., 2010; Hoyle et al., 2001), the use of MLMs is advised to analyse the data. However, when the clustered data are negatively correlated, the dissimilarity in the clustered data is usually ignored, despite the negative effects of ignoring a negative ICC. Even though others –such as Kenny et al. (2002) and Pryseley et al. (2011)– already drew attention to this phenomenon of dissimilarity, it is obvious that the negative counterpart is less well understood. Furthermore, a more pragmatic reason is that until recently, the tools to study negatively correlated data is lacking (although there are of course exceptions, Molenberghs & Verbeke, 2011; Verbeke & Molenberghs, 2003). The opinion is that this risk of ignoring a non-zero ICC is currently even greater under negative clustering effects, as MLMs cannot assess negative clustering effects, and the effects of negative clustering effects appear to be less well-known by researchers.

viii

Go Beyond Sample Size Restrictions

To obtain stable parameter estimates for the MLM, the sample size needs to be sufficient for the different levels of the model. For the one-way random effects model, a sufficient number of clusters is needed to estimate the variability across

[3]Nielsen et al. (2020) is chapter 7 of this thesis.

groups. For a multi-way random effects model, for each clustering type a sufficient number of clusters are needed to obtain a stable random effect variance estimate. Maas and Hox (2005) reported that a small sample size at level two can lead to biased estimates of the second-level standard errors. A small number of level-two groups can lead to a zero level-two variance estimate, indicating that there is simply not enough information. The Bayesian approach can introduce a prior to by-pass this problem. However, a prior distribution can force the variance estimate to be positive. This can highly depend on the specified prior and might not represent correctly the variation across clusters in the population. Furthermore, the motivation for doing a multilevel analysis is that the sample size within each cluster is less than overwhelming. Then, the cluster-level variance is used as a weight to reduce the error in the cluster-specific estimates by pooling information across clusters. However, the shrinkage in the cluster-specific estimates might be less than desired, when the cluster-level variance is overestimated.

In the BCSM, the dependence structure is modelled through a common covariance parameter for the clustered observations. This reduces the sample size restrictions for the BCSM compared to the MLM. Furthermore, a prior for a covariance parameter is not restricted to positive values. The BCSM can be applied to a two-stage (or multi-stage) sample, where clusters are sampled independently, and subsequently observations within each cluster are independently sampled. However, by modelling directly the covariance among clustered observations, the BCSM also applies to a stratified sample in which independent samples are drawn for the considered clusters.

We will demonstrate in our simulation study that even for two clusters stable covariance parameter estimates can be obtained. Furthermore, the BCSM will prove to be very useful for analysing small data sets. Under the BCSM, the practical definition of what is considered a small sample size changes considerably. The BCSM in Equation (8.4) does not contain any cluster-specific parameters, although cluster-specific estimates can be obtained from fitted residuals. As a result, cluster-level variance estimates are not needed to shrink cluster-specific parameters. This avoids the issue of estimating the variability across clusters, and to use those estimates to reduce errors in the clusters-specific parameter estimates by shrinking them. This makes the BCSM much more suitable for small sample sizes than the MLM. In the BCSM, it is not needed to explicitly model variability across clusters and to estimate any cluster-specific (i.e. random effect) parameters. Furthermore, due to the Bayesian modelling approach, it is also not necessary to rely on large sample theory to make statistical inferences.

Model Complexity

The BCSM represents a far more parsimonious way to model a dependence structure than the random effects approach in MLM. Under the BCSM, the number of covariance parameters to model the dependence structure does not depend on the sample size. This in contrast to the MLM, where the required number of random effect parameters depends on the number of clusters. Indeed, increasing the number of clusters does not affect the complexity of the BCSM, where the MLM becomes more complex. Furthermore, for each additional type of clustering, the dimensionality of the MLM increases and requires an additional set of random effect parameters, where the BCSM requires just one additional covariance parameter.

The BCSM can even model a dependence structure implied by non-identified random effects. For instance, assume pre-intervention and post-intervention data of persons, and let α_i denote the person-specific random effect for the post-measurement of person i, which is normally distributed with variance τ_a. The cluster size is $n = 1$ (i.e. each person has one post-measurement), which makes it impossible to estimate the random effect α_i and the variance τ_a. Under the BCSM, the parameter τ_a is identified and can be estimated, which provides information about the dependence of the post-intervention measurements. The BCSM approach is straightforward and elegant: the covariance matrix has a common error variance σ^2 for the pre-intervention measurements and a variance component τ_a is added to the common error variance for the post-intervention measurements. The heteroscedastic error variances of the covariance matrix are identified and can be motivated by the (unidentified) random effect α_i. Thus, under the BCSM, the dependence structure of a random interaction effect can be estimated from clusters which only have one observation.

Unbiased Estimator: Include the Entire Parameter Space

Common maximum likelihood (ML) and Bayesian estimation methods restrict the random effect variance estimate to be positive. Bayesian methods use a prior which assigns a positive density to non-negative values; ML methods usually restrict the variance estimate to be positive, although negative variance estimates are possible (see below). This leads to biased parameter estimates. We show here that the random intercept model, Equation (8.1), gives support to data sets for which the ML estimate is negative. As a result, when not allowing negative variance estimates, the ML estimator is biased, since the negative parameter space of the sampling distribution of the estimator is ignored. This also holds for the restricted maximum likelihood estimator and for (un)balanced designs. In the

BCSM, the prior for the covariance parameter includes the negative parameter space, for all values for which the covariance matrix is positive definite. As a result, an estimator for the covariance parameter under the BCSM is not biased, since the entire parameter space is taken into account.

A negative ML estimate of the random effect variance has received attention (El Leithy et al., 2016; Kenny et al., 2002; Loeys & Molenberghs, 2013; Molenberghs & Verbeke, 2007, 2011; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003), partly due to the embarrassment of obtaining a negative estimate for a parameter which by definition is non-negative (Searle et al., 1992, p.60). For the random intercept model in Equation (8.1), it can be easily seen that the ML estimate for the random intercept variance τ can be negative depending on the observed between-cluster and (within-cluster) error sum of squares. For balanced groups, the two sums of squares are considered to estimate the covariance component τ,

$$\begin{aligned} SS_A &= \sum_{i=1}^{a} n\left(\overline{y}_i - \overline{y}\right)^2, \\ SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_j\right)^2. \end{aligned}$$

The sum of squares SS_A/a has expected value $n\tau + \sigma^2$. It follows that,

$$\begin{aligned} \hat{\tau} &= \frac{\frac{SS_A}{a} - \frac{SS_E}{n(a-1)}}{n} \\ &= \frac{\frac{SS_A}{a} - MSE}{n}, \end{aligned}$$

using the MSE as an estimator for σ^2. The estimate for τ is negative when $MSE > SS_A/a$. The negative estimates are neglected or referred to as statistically incorrect, restricting τ to be positive, $0 < \tau \leq \infty$. However, the ML estimate is not necessarily in this parameter space, which occurs with probability $P(MSE > SS_A/a)$. As described by McCulloch et al. (2008), the ML estimator has two possible outcomes

$$\hat{\tau} = \begin{cases} \hat{\tau} & \text{if} \quad SS_A/a \geq MSE \\ 0 & \text{if} \quad SS_A/a < MSE. \end{cases}$$

The estimate of the variance is restricted to be zero, when the data gives support to a negative estimate. Of course this makes sense, since τ represents a variance component. However, for $\tau < 0$ there is cluster dissimilarity, which will be interpreted incorrectly as cluster similarity when τ is restricted to be positive.

Solving Boundary Issues

In the MLM, the random effect variance is restricted to be greater or equal to zero. This value of zero is a lower bound but also of specific interest. A random effect variance of zero implies that the groups do not differ, where a positive variance implies that the groups differ. It is well-known that classical test procedures such as the likelihood-ratio test can break down and leads to inconsistent testing, when testing if a parameter lies on the boundary of the parameter space.

In the Bayesian framework, test and estimation methods depend on the specified prior distributions. Specifying a prior for a random effect variance is a complicated task, since the point zero is a boundary value. The popular conjugate inverse-gamma prior only gives support to positive values. The exact specification of the prior depends on the hyper parameter values. When the variance is near zero the hyper parameters need to be close to zero. Most often the default inverse-gamma prior is sharply peaked near zero to give support to variance values near zero. Thus, an objective (non-informative) prior specification is not possible without knowing the true parameter value. Otherwise stated, the posterior distribution is sensitive to the hyper parameter values of the inverse-gamma distribution. Gelman (2006) recommended different classes of priors such as the half-t family of prior distributions, to improve the behaviour of the prior near zero. However, the priors are not completely objective and, in general, place too much mass on higher variance values when the true value is close to zero. This phenomenon is shown in our simulation study.

Under the BCSM, the value $\tau = 0$ is not a lower bound. Therefore, a noninformative prior can be specified for those parameter values that ensure a positive-definite covariance matrix. Following Fox et al. (2017), a truncated shifted inverse-gamma prior can be specified that allows the parameter space to cover also negative values while enforcing sufficient rules for the positive definiteness of the covariance matrix. These priors are not sharply peaked near zero such as the default inverse-gamma priors but remain uninformative about the presence of negative, positive, or zero correlation. In addition, with the shifted inverse-gamma prior, more accurate estimates of a very small random-effect variance can be obtained by avoiding too much prior support for higher parameter values.

Parameter Estimation for the BCSM

A general technique is proposed to estimate the model parameters of the BCSM. The estimation method is based on a Gibbs sampler (Markov chain Monte Carlo,

MCMC), where the variance components of the BCSM can be directly sampled from their conditional posterior distributions. The posterior distribution of each variance component can be analytically derived from which parameter values can be directly sampled. This technique is based on a balanced design, which means that the number of observations is equal across the same type of clustering. Although the BCSM is by no means limited to balanced designs alone, the extension to unbalanced designs is beyond the scope of our current study.

One-way Classification

Three steps can be defined to construct the MCMC algorithm for the BCSM for the one-way classification in Equation (8.4). In step 1, the expected within-sum of squares (SS_E) is derived to construct the posterior distribution of the variance parameter σ^2. In a similar method, in step 2, the expected between-sum of squares (SS_A) is derived. In step 3, a shift-parameter is introduced for the result of step 2, to obtain the posterior distribution of the covariance parameter τ.

The posterior distributions of the variance components σ^2 and τ are derived. In this model, the total sum of squares (SS_T) is partitioned in a between- and within-sum of squares, referred to as SS_E and SS_A (type-A clustering), respectively,

$$\begin{aligned} SS_T &= SS_A + SS_E \\ \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_{..}\right)^2 &= \sum_{i=1}^{a} n\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_{i.}\right)^2, \end{aligned} \tag{8.8}$$

where $\overline{y}_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}/(na)$ and $\overline{y}_{i.} = \sum_{j=1}^{n} y_{ij}/n$. The part of the likelihood that includes the general mean is excluded. This follows from partitioning the likelihood;

$$p\left(\mathbf{y} \mid \mu, \sigma^2, \tau\right) = p\left(\mu \mid \overline{y}_{..}\right) p\left(\sigma^2, \tau \mid SS_E, SS_A\right),$$

see, for instance, McCulloch et al. (2008). As they follow directly from standard Bayesian linear regression theory (Gelman et al., 2013), the posterior distributions of fixed effect parameters are not discussed.

The conditional model in Equation (8.1) in which observations are conditionally independently distributed given cluster-specific parameters, is used to find the model expressions for the cluster and sample-averaged observations. It

follows that

$$\overline{y}_{i.} = \mu + \alpha_i + \overline{e}_{i.} \tag{8.9}$$

$$\overline{y}_{..} = \mu + \overline{\alpha}_{.} + \overline{e}_{..}, \tag{8.10}$$

where $\overline{e}_{i.} \sim N(0, \sigma^2/n)$ and $\overline{e}_{..} \sim N(0, \sigma^2/(na))$. The expressions are used to obtain the expected sum of squares under the model.

Step 1 is carried out. Therefore, the expected value of the SS_E is derived by integrating the model expression for the cluster mean (Equation 8.9):

$$\begin{aligned}
E(SS_E) &= E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{i.}\right)^2\right) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n} E\left(\left(\mu + \alpha_i + e_{ij}\right) - \left(\mu + \alpha_i + \overline{e}_{i.}\right)\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n} E\left(e_{ij} - \overline{e}_{i.}\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n} E\left(e_{ij}^2\right) - E\left(\overline{e}_{i.}^2\right) \\
&= an\left(\sigma^2 - \frac{\sigma^2}{n}\right) \\
&= a(n-1)\sigma^2,
\end{aligned} \tag{8.11}$$

where in the extraction of the binomial product the inner product cancels (from the second to the third expression, and the third to the fourth expression), since the expected value of each error term is zero. For a balanced design and pairwise independent SS_E components, the SS_E divided by their expected value is (central) chi-square distributed (Searle, 1971, p.174).

Assume an inverse-gamma prior for σ^2, $\sigma^2 \sim IG(g_1/2, g_2/2)$. Then, the posterior distribution of the σ^2 is an inverse-gamma distribution with SS_E as the sufficient statistic (Gelman et al., 2013),

$$p\left(\sigma^2 \mid \mathbf{y}\right) \propto \left(\sigma^2\right)^{-((a(n-1)+g1)/2+1)} \exp\left(-\frac{(SS_E + g_2)/2}{\sigma^2}\right) \tag{8.12}$$

with shape parameter $(g_1 + a(n-1))/2$ and scale parameter $(SS_E + g_2)/2$. For $g_1 = 0$ and $g_2 = 0$ the uninformative reference prior is specified for σ^2.

In **step 2**, a similar procedure is followed for the covariance parameter τ.

Consider the between sum of squares SS_A,

$$\begin{aligned} E(SS_A) &= E\left(n\sum_{i=1}^{a}(\overline{y}_{i.}-\overline{y}_{..})^2\right) \\ &= n\sum_{i=1}^{a}E\left((\mu+\alpha_i+\overline{e}_{i.})-(\mu+\overline{\alpha}_{.}+\overline{e}_{..})\right)^2 \\ &= n\sum_{i=1}^{a}E\left((\alpha_i-\overline{\alpha}_{.})+(\overline{e}_{i.}-\overline{e}_{..})\right)^2 \\ &= n\sum_{i=1}^{a}E\left(\alpha_i-\overline{\alpha}_{.}\right)^2+E\left(\overline{e}_{i.}-\overline{e}_{..}\right)^2 \\ &= n\sum_{i=1}^{a}\left(E(\alpha_i^2)-E(\overline{\alpha}_{.}^2)\right)+\left(E\left(\overline{e}_{i.}^2\right)-E\left(\overline{e}_{..}^2\right)\right) \\ &= an\left(\left(\tau-\frac{\tau}{a}\right)+\left(\frac{\sigma^2}{n}-\frac{\sigma^2}{an}\right)\right) \\ &= (a-1)\left(n\tau+\sigma^2\right), \end{aligned} \tag{8.13}$$

where the inner product of the binomial products is again zero, since the expected error terms are equal to zero. The SS_A/n is considered as the sufficient statistic for the term $\lambda = \tau + \sigma^2/n$, which has an inverse-gamma distribution. The λ is restricted to be positive, which means that $\tau > -\sigma^2/n$ with $\sigma^2 > 0$.

In **step 3**, the shift parameter is introduced, which is the term σ^2/n, and allows the τ to take on negative values. This restriction on the parameter space of τ can be defined in the noninformative prior for τ;

$$p\left(\tau \mid \sigma^2\right) \propto \left(\tau+\sigma^2/n\right)^{-1}, \tag{8.14}$$

since it restricts the τ to be greater than $-\sigma^2/n$ with $\lambda = \tau + \sigma^2/n$ restricted to be greater than zero. Following Fox et al. (2017) and Klotzke and Fox (2019a), the posterior distribution of τ is referred to as a shifted inverse-gamma distribution

$$p\left(\tau \mid \mathbf{y}, \sigma^2\right) \propto \left(\tau+\sigma^2/n\right)^{-((a-1)/2+1)} \exp\left(-\frac{(SS_A/n)/2}{\tau+\sigma^2/n}\right).$$

It can also be shown that for all τ values above this lower bound the covariance matrix in Equation (8.2) is positive definite (Fox et al., 2017). Parameter values from this shifted inverse gamma distribution can be obtained by sampling $\lambda^{(m)}$ from an inverse-gamma distribution with $(a-1)$ degrees of freedom and scale parameter SS_A/n in iteration m. Then, a sampled value for τ is obtained by subtracting the sampled value for σ^2, $(\lambda^{(m)} - \sigma^2/n)$.

Two-way Classification

This procedure to derive the posterior distributions of the variance components can be extended to covariance parameters for other cross-classified and/or nested factors. Without giving a general description, the two-way nested classification model in Equation (8.6) is considered to illustrate the procedure for two types of clustering (referred to as type A and type B). Again three steps can be defined, where step 1 is similar to the step 1 for the one-way classification. Then, step 2a (obtain expected between sum of squares) and 3a (derive shift parameter) are defined to obtain the posterior distribution of parameter τ_a for the clustering of type A. Analogously, step 2b and 3b are defined for the τ_b for the clustering of type B.

The total sum of squares is partitioned in three components, the total sum of squares (SS_T), a sum of squares SS_A (cluster A), a sum of squares SS_B (cluster B) and a within-sum of squares (SS_E):

$$\begin{aligned} SS_T &= SS_A + SS_B + SS_e \\ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\overline{y}_{...})^2 &= \sum_{i=1}^{a} nb\,(\overline{y}_{i..}-\overline{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b} n\left(\overline{y}_{ij.}-\overline{y}_{i.}\right)^2 \\ &\quad + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk}-\overline{y}_{ij.}\right)^2. \end{aligned} \tag{8.15}$$

Step 1: the expected value of the SS_E is derived,

$$\begin{aligned} E(SS_E) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} E\left(y_{ijk}-\overline{y}_{ij.}\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} E\left(e_{ijk}-\overline{e}_{ij.}\right)^2 \\ &= abn\left(\sigma^2-\frac{\sigma^2}{n}\right) = ab(n-1)\sigma^2 \end{aligned}$$

It follows that the posterior distribution of the variance parameter σ^2 is an inverse gamma distribution, with the $SS_E/2$ as the scale parameter. The variance parameter has an inverse-gamma distribution with shape parameter $(g_1+ab(n-1))/2$ and scale parameter $(g_2+SS_E)/2$.

Then in **step 2b**, the posterior distribution of the covariance parameter τ_b is derived by determining the expected value of the SS_B, which is the sufficient

statistic. It follows that,

$$\begin{aligned} E(SS_B) &= E\left(\sum_{i=1}^{a}\sum_{j=1}^{b} n\left(\overline{y}_{ij.} - \overline{y}_{i..}\right)^2\right) \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b} nE\left((\mu + \alpha_i + \beta_{ij} + \overline{e}_{ij.}) - (\mu + \alpha_i + \overline{\beta}_{i.} + \overline{e}_{i..})\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b} nE\left(\beta_{ij} - \overline{\beta}_{i.}\right)^2 + nE\left(\overline{e}_{ij.} - \overline{e}_{i..}\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b} n\left(\tau_b - \frac{\tau_b}{b}\right) + n\left(\frac{\sigma^2}{n} - \frac{\sigma^2}{nb}\right) \\ &= a(b-1)\left(n\tau_b + \sigma^2\right). \end{aligned} \tag{8.16}$$

The prior for the parameter τ_b is defined as

$$p\left(\tau_b \mid \sigma^2\right) \propto \left(\tau_b + \sigma^2/n\right)^{-1}, \tag{8.17}$$

which allows τ_b to be negative but greater than $-\sigma^2/n$. **Step 3b**: The posterior distribution for τ_b is a shifted inverse-gamma distribution with shape parameter $a(b-1)/2$, scale parameter SS_B/n and shift parameter σ^2/n.

Step 2a: the posterior distribution of the covariance parameter τ_a can be obtained in the same way, by considering the expected sum of squares of SS_A,

$$\begin{aligned} E\left(SS_A\right) &= bn\sum_{i=1}^{a} E\left((\mu + \alpha_i + \overline{\beta}_{i.} + \overline{e}_{i..}) - (\mu + \overline{\alpha}_{.} + \overline{\beta}_{..} + \overline{e}_{...})\right)^2 \\ &= bn\sum_{i=1}^{a} E\left(\alpha_i - \overline{\alpha}_{.}\right)^2 + E\left(\overline{\beta}_{i.} - \overline{\beta}_{..}\right)^2 + E\left(\overline{e}_{i..} - \overline{e}_{...}\right)^2 \\ &= bna\left(\left(\tau_a - \frac{\tau_a}{a}\right) + \left(\frac{\tau_b}{b} - \frac{\tau_b}{ab}\right) + \left(\frac{\sigma^2}{bn} - \frac{\sigma^2}{abn}\right)\right) \\ &= (a-1)\left(bn\tau_a + n\tau_b + \sigma^2\right). \end{aligned} \tag{8.18}$$

The $SS_A/(bn)$ is the sufficient statistic for the τ_a, then the prior for τ_a equals

$$p\left(\tau_a \mid \tau_b, \sigma^2\right) \propto \left(\tau_a + \left(\tau_b/b + \sigma^2/(bn)\right)\right)^{-1}. \tag{8.19}$$

Step 3a: it follows that the posterior distribution of τ_a is shifted inverse-gamma with shape parameter $(a-1)$, scale parameter $SS_A/(bn)$, and shift parameter $\tau_b/b + \sigma^2/(bn)$. The τ_a is restricted to be greater than $-(\tau_b/b + \sigma^2/(bn))$, where $\tau_b > -\sigma^2/(n)$.

Multi-way Classification

In a more general description, for a balanced design a Gibbs sampling procedure can be defined for any multi-way classification model, where different types of clustering group the continuous data. The (lower-level) variance parameter has an inverse-gamma posterior distribution, where the SS_E is the sufficient statistic. Each covariance parameter has a shifted inverse-gamma distribution, which is constructed from the sum of squares representing the corresponding sufficient statistic. The parameter space of the variance components covers those negative values that still lead to a positive definite covariance matrix. In the Gibbs sampling algorithm, the variance components can be iteratively sampled from their posterior distributions, which leads to a very fast and efficient sampling method.

The MCMC algorithm is easily extended when including a sampling step for fixed effect parameters. Consider the BCSM in Equation (8.4), and let $\mu = \mathbf{X}_i\boldsymbol{\beta}_f$. The covariance matrix $\boldsymbol{\Sigma}$ has two parameters σ^2 and τ, and the inverse of the covariance matrix is known (Searle et al., 1992). When assuming an uniform prior, the posterior distribution for $\boldsymbol{\beta}_f$ is normal with variance and mean

$$\begin{aligned} Var\left(\boldsymbol{\beta}_f \mid \mathbf{y}, \boldsymbol{\Sigma}\right) &= \boldsymbol{\Omega} = \left(\mathbf{X}^t \left(\mathbf{I}_a \otimes \boldsymbol{\Sigma}^{-1}\right) \mathbf{X}\right)^{-1}, \\ E\left(\boldsymbol{\beta}_f \mid \mathbf{y}, \boldsymbol{\Sigma}\right) &= \boldsymbol{\Omega}\mathbf{X}^t \left(\mathbf{I}_a \otimes \boldsymbol{\Sigma}^{-1}\right) \mathbf{y}, \end{aligned}$$

respectively.

Simulation Study

The BCSM estimation method was investigated for small variance components close to the lower bound of zero, for few clusters, few observations for each cluster, and even for negative cluster dependences. Data were simulated under a random intercept model, where the residual variance was equal to $\sigma_e^2 = 5, 1, 0.5, 0.1, 0.01$ and the random intercept variance was equal to $\tau = 5, 1, 0.5, 0.1, 0.01$. The general mean was simulated from a standard normal distribution. The number of clusters was equal to $a = 50, 25, 10, 5$, and the number of observations per cluster $n = 20, 10, 5, 2$. All conditions were crossed with each other resulting in 400 simulation conditions. For each condition, 1,000 data replications were made according to the random intercept model defined in Equation (8.1), and they are referred to as the conditional data. The conditional data was analysed with LME4, which produced (restricted) maximum likelihood (REML) estimates for the variance components. Furthermore, an MCMC estimation method was used (JAGS), with (vague) inverse-gamma priors for the variance components

(shape and scale parameter equal to .01), and a noninformative normal distribution for the general mean. The median of posterior distribution was used as a point estimator for the variance components, since these distribution were often asymmetric.

In the BCSM, the parameter τ is a covariance parameter, which can also be negative. Therefore, data was also generated with τ negative but just above the lowerbound; $L_b = -\sigma^2/n + 10^{-4}$, which assured that the covariance matrix was positive definite. Data was simulated under the BCSM, as defined in Equation (8.4), for the same conditions as described for the random intercept model. The condition $\tau = L_b$ was added for each combination of a, n, and σ^2. This led to a total of 480 conditions. For each condition, 1,000 data sets were generated under the BCSM, and referred to as marginal data. The data was analysed with the BCSM, LME4, and JAGS. The main interest was the estimation of the (co)variance component, τ.

The RMSE and 95% coverage rate (CR) was used as a criterion to evaluate the estimation results. The estimated CR represented the proportion that the true parameter value was covered by the 95% credible interval (CI) across the 1,000 data replications and should be around the advocated 95%. The 95% CIs were computed using the MCMC samples. The MCMC algorithms for the BCSM and random intercept model in JAGS were ran for 10,000 iterations, while using 500 iterations as the burn-in period. The MCMC samples showed good convergence in each condition, which was inspected using the MCMC convergence tools in the `coda` R-package. The effective sample size was around 90% for both MCMC methods for all model parameters.

Negative cluster dependence was simulated in the marginal data under the BCSM (Equation 8.4 and 8.5), since this could not be done with the random intercept model. The results discussed below concerning a true negative cluster correlation concerns the marginal data (generated under the BCSM). The conditional data generated under the random intercept model was used to evaluate the performance of LME4 and JAGS, when the true value of τ was positive. The marginal data generated under the BCSM was used to evaluate the performance of the BCSM for all values of τ.

viii

Negative within-cluster correlation

In Table 8.1, it can be seen that the negative cluster dependence, τ, was accurately estimated under the BCSM for all conditions. Table 8.1 also shows that biased estimates were obtained with LME4 and JAGS. The negative lower bound L_b for τ differs across conditions. The random intercept model cannot describe neg-

Table 8.1: MSE and 95% coverage rates of the lower bound (L_b) of $\hat{\tau}$.

a	L_b	$n = 20$	$n = 10$	$n = 5$	$n = 2$
			LME4		
50	−0.25	0.10	0.33	1.13	6.43
25	−0.50	0.15	0.40	1.21	6.52
10	−1.00	0.28	0.71	1.76	6.79
5	−2.50	0.50	1.08	2.39	8.36
			JAGS		
50	−0.25	0.08 (.00)	0.31 (.00)	1.18 (.00)	7.28 (.00)
25	−0.50	0.09 (.00)	0.33 (.00)	1.25 (.00)	7.71 (.00)
10	−1.00	0.10 (.00)	0.37 (.00)	1.38 (.00)	8.39 (.00)
5	−2.50	0.13 (.00)	0.44 (.00)	1.57 (.00)	9.61 (.00)
			BCSM		
50	−0.25	0.00 (.95)	0.00 (.95)	0.00 (.95)	0.00 (.95)
25	−0.50	0.00 (.95)	0.00 (.95)	0.00 (.95)	0.00 (.94)
10	−1.00	0.00 (.95)	0.00 (.93)	0.00 (.93)	0.00 (.95)
5	−2.50	0.00 (.95)	0.00 (.94)	0.00 (.95)	0.00 (.93)

ative cluster dependence, and estimation results under LME4 and JAGS show high RMSEs and incorrect CIs for the cluster dependence. Table 8.1 shows that accurate estimates were obtained under the BCSM even in the extreme condition of just five groups with each two observations. The BCSM still performed good with a total of only ten observations. In that condition the RMSE is still less than .00 and the CR is around 93%.

Small variance component

The estimation methods, referred to as LME4 (REML), JAGS, and BCSM, performed comparable in the conditions with sufficient data to estimate the parameters. When considering the τ estimates, the parameter estimates are alike when the τ is positive (but not close to zero), and there is sufficient data with respect to the number of groups and the number of observations per group. In Figure 8.2, the bottom plot shows the estimates for τ averaged across 1,000 replications under LME4, JAGS, and the BCSM in the condition with $a = 50$ groups and $n = 5$ observations per group with the residual variance σ^2 and τ varying across the different specified levels. For each τ value, data were generated with five different residual variances ranging from 5 to .01. In total 30 estimates for τ are plotted for each method. It can be seen that the estimates are close to the true value. In the standard situations, Figure 8.2 shows that the BCSM performs on

par with the standard estimation methods.

However, when the true value is almost zero or close to zero, then the BCSM outperforms the other methods. In that case the BCSM still provides accurate estimates for τ for all values of the residual variance. The upper plot shows an extreme condition with just five groups with each two observations. It can be seen that the BCSM estimates are still close to the true value, but it tends to underestimate a true τ of five. With JAGS, the τ is overestimated in more cases when the true value is close to zero or negative. The inverse-gamma prior for τ led to an overestimation of the true value, although the hyper parameter values were .01. The REML results with LME4 also overestimated the true value, when it was negative. Furthermore, in the situation with a lack of prior information and poor data information, the REML estimates were much higher than the true values.

MSE

When considering the MSEs for the τ estimates, JAGS, LME4 and BCSM performed comparably good when there is sufficient data information. However, for true negative values of τ, the BCSM outperformed the other methods. Furthermore, the MSEs of the REML estimates are much higher, when the residual variance is equal to five. In the small sample conditions, the BCSM outperformed both other methods. The MSEs under JAGS and LME4 are higher when the true value is close to zero or negative. When the residual variance is five, the MSEs under LME4 are also much higher than for the other methods. For all the considered conditions, the MSEs under the BCSM are close to zero.

Coverage

Finally, the 95% CRs were computed under JAGS and the BCSM method. Confidence intervals for the variance components were not computed under maximum likelihood estimation (LME4), since this led to numerical problems and invalid CIs (e.g. using bootstrap function in LME4 to compute CIs). The estimated CRs for JAGS were zero, when the true value was negative. The inverse-gamma prior for τ restricted the posterior distribution of τ to only cover positive values, which led to incorrect CRs. Under JAGS, the CRs were too large and close to one, when the true value of τ was positive but close to zero, and when there was not much data. In the situation without much data information, the posterior was more stretched by the prior which led to wider credible regions than expected under the data replications.

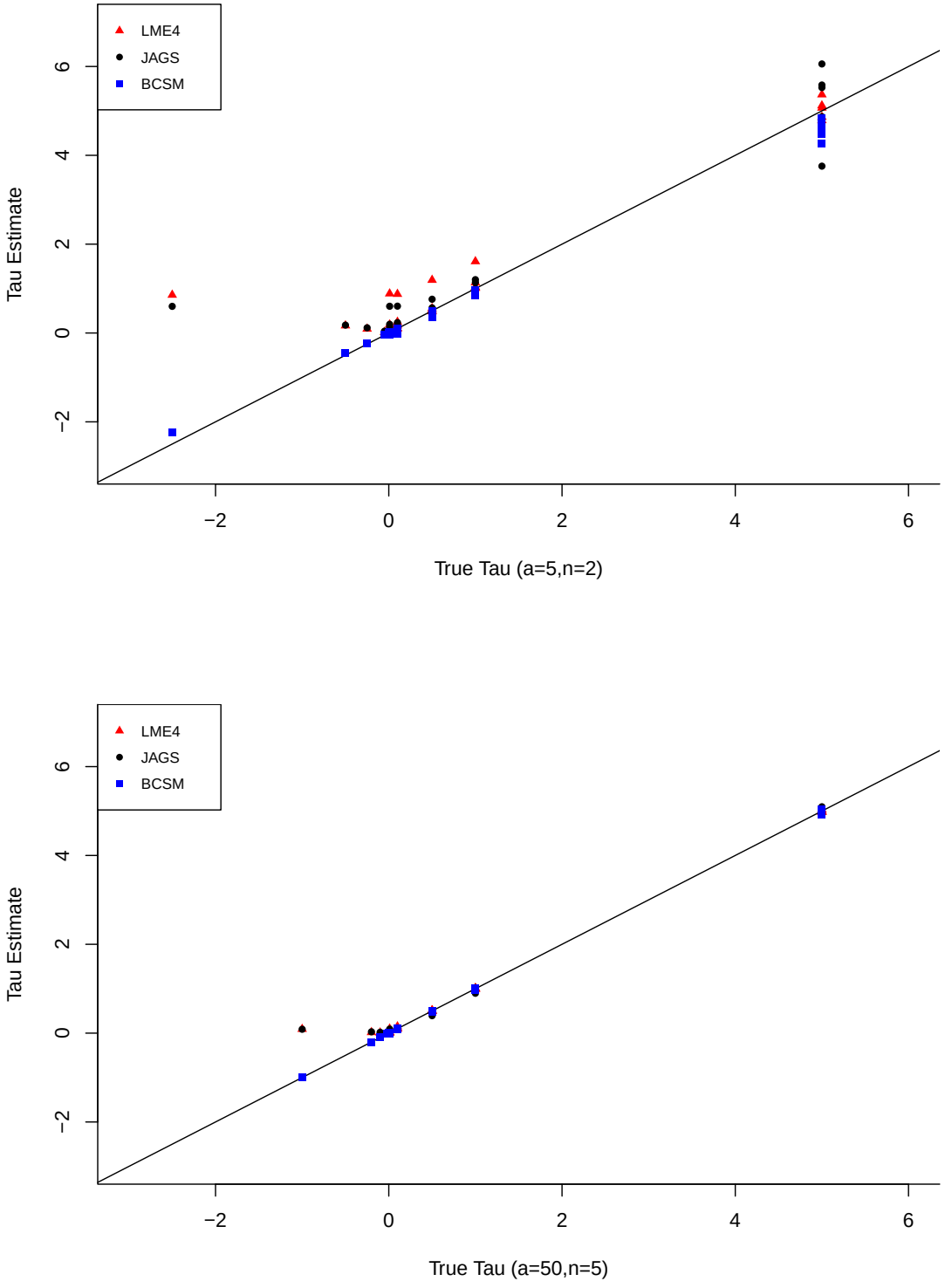


Figure 8.2: Averaged parameter estimates for τ across 1,000 data replications under LME4, JAGS, and the BCSM.

The data were generated for fixed values for τ, which means that the prior variance was not included in the data replications under the BCSM and random

intercept model. This led to an overestimation of the 95% CR in the extreme data conditions, when the prior variance influenced the width of the credible regions. For instance, for JAGS, with $\sigma^2 = 1$ and $\tau = .01$ the estimated 95% CRs were around $.98-1$, but mostly one for all considered samples sizes. In those conditions, the estimated CRs were still around 95% for the BCSM, except when $a = 5$ and $n = 2$ the CR was also one. In general, accurate 95% CRs were computed for the BCSM. Only in the extreme conditions, when the number of clusters was five, and the residual variance and τ were both small (i.e., $\sigma^2 = .1$, $\tau = .01$), the estimated CRs were close to one for the BCSM. For JAGS, the 95% CRs were more often overestimated also for other values of the residual variance and also for conditions with more than five groups. The inverse-gamma prior for τ in JAGS influenced more the 95% credible regions and gave more weight to higher values than the shifted-inverse gamma prior in the BCSM. This is a typical issue for inverse-gamma priors for variance components in hierarchical models (Gelman, 2006). In the BCSM, the shifted-inverse gamma prior performed better simply by extending the parameter space to include negative values.

Personalized Treatment in E-mail Counselling

The effectiveness of BCSM is demonstrated for a real-data example. Lamers et al. (2015) examined whether a combination of a self-help intervention with narrative therapy is effective in alleviating symptoms of depression and anxiety. The treatment consisted out of two conditions: the *auto-biographic* and the *expressive* writing condition (AW and EW, respectively). The AW condition was a life-review self-help intervention that consisted of homework assignments, divided over modules that had to be completed over the course of ten weeks. Clients communicated about their progress with trained counsellors through a weekly e-mail interaction. The EW intervention was based on the method of expressive writing. The method consisted of daily writing about emotional experiences, for $15 - 30$ minutes on $3 - 4$ consecutive days during one week. Lamers et al. (2015) used a repeated measures ANOVA and found that depressive symptoms indeed declined, but did not find a difference between the AW and EW condition (in comparison with a waiting list control group). Smink, Fox, et al. (2019) adopted a multilevel approach with client as a random effect, and also did not find a significant difference between treatments.

The BCSM was used to identify individual variability in treatment effects and to identify those who benefitted from the treatment, since a significant main treatment effect could not be found. Furthermore, the object was to investigate

the effect of the counsellor and how they contributed to the treatment of the clients. Several clients in different treatment arms were treated by the same counsellor, and negative clustering effects were expected since individualized treatments were given by each counselor. Let i denote the index for the counsellor and j the client. Each client was measured at a pre- and a post-intervention occasion, which resulted in a y_{ij1} and y_{ij2} score, respectively. Scores of clients who were treated by the same counsellor (i.e. counsellor i) were assumed to be clustered, and we also assumed that scores coming from the same client (i.e. client j) were clustered. Let factor variable α_i represent the counselling effect, and nested factor variable $\beta_{(i)j}$ the client effect. This leads to a two-way nested factor model for the pre- and post-intervention scores presented in Equation (8.6). In the corresponding BCSM, the covariance structure implied by the two factor variables was directly modelled, allowing for the occurrence of potentially negative cluster correlations.

Measuring client, counsellor and individual treatment effects

We first fitted a linear regression model, denoted as LM M0, which assumed independently distributed errors. Then, two BCSMs were considered, to which we refer as M1 and M2, which had the same mean term as the LM M0. For all three models, the intercept β_0 represented the average score at the pre-intervention for clients in the AW condition. The treatment variable was dummy-coded (with a *one* for clients in condition EW, and a *zero* for those in condition AW). The main effect of treatment, represented by β_1, was included to correct for any pre-intervention differences between the two treatment groups. The β_2 represented the average contribution of the post-intervention in comparison to the pre-intervention score, where indicator variable $Post_{ij}$ was also dummy-coded (with a *one* for the post-intervention scores, and *zero* for the pre-intervention scores). An interaction variable Z with effect β_3 was dummy coded, where a *one* represented the interaction between the post-intervention measurement of clients in condition EW.

BCSM M1 and M2 assumed dependence among scores from clients assigned to the same counsellor, and M2 also assumed a dependence among pre- post-intervention scores from the same client. To better understand the factor structure represented in the covariance structure of BCSM M2, consider the (conditional) MLM with random effects for the counsellor and the client;

$$\begin{aligned} y_{ijl} &= \beta_0 + \beta_1 Treatment_{ijl} + \beta_2 Post_{ijl} + \beta_3 Z_{ijl} + \alpha_i + \beta_{(i)j} + e_{ijl}, \\ \alpha_i &\sim N(0, \tau_a) \text{ (Counsellor)} \end{aligned}$$

$$\begin{aligned} \beta_{(i)j} &\sim N(0, \tau_b) \text{ (Client)} \\ e_{ijl} &\sim N(0, \sigma^2), \end{aligned}$$

where $l = 1, 2$ indicates a pre-intervention or post-intervention observation, respectively. The MLM cannot detect negative clustering effects, and it needs 90 random client parameters and five random counsellor parameters to model the dependence structure. This makes it unsuitable for the small data set. Therefore, the dependence structure is directly modeled, which leads to the following BCSM:

$$\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{Treatment}_i + \beta_2 \mathbf{Post}_i + \beta_3 \mathbf{Z}_i + \mathbf{E}_i, \tag{8.20}$$

where the $\mathbf{E}_i$ is (multivariate) normally distributed. The three models –LM Mo, BCSM M1, and BCSM M2– can be represented by the model in Equation (8.20), but each model has its specific (structured) covariance matrix $\mathbf{\Sigma}$. For model LM Mo, the covariance matrix $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$ represents independently distributed errors. For BCSM M1, a one-way clustering is assumed represented by the covariance matrix $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n + \tau_a \mathbf{J}_n$. For BCSM M2, the covariance matrix $\mathbf{\Sigma}$ is given in Equation (8.7). The covariance matrix of M1 represents a one-way clustering with τ_a the covariance of scores of those treated by the same counsellor. The covariance matrix of BCSM M2 also includes a component τ_b, which represents the covariance of scores of the same client.

For the BCSM models M1 and M2, an MCMC algorithm with 20,000 iterations (with a burn-in of 1,000 iterations) was used to compute the parameter estimates. The parameter estimates of BCSM M1 and M2 are given in Table 8.2. The BCSM M2, with a two-nested dependence structure, contained four regression parameters and three (co)variance parameters. This makes the BCSM particularly useful for small data sets. A trimmed mean estimator was used for the covariance components, where 10% of the outlying values were ignored to obtain more robust posterior mean estimates. The posterior standard deviations were estimated using all sampled values. The parameter estimates of model Mo were obtained using the `lm`-function in `R`.

The parameter estimates of the regression effects did not differ much for the different models. The adjusted R^2 was around .91 under model LM Mo. It can be seen that on average on the post-intervention clients scored four points lower than on the pretest, showing that depressive symptoms indeed declined. There were no significant differences between the two treatment groups on the pre-intervention. The interaction effect β_3 was around -1.36, showing that those in the EW condition scored on average lower than those in the AW condition at the

Table 8.2: The e-mail-counselling study from Lamers et al. (2015): A BCSM analysis of the pre- and post-therapeutic data.

		LM ($\hat{M}$, $S.E.$)	BCSM ($\hat{M}$, SD)		
		M0	M1	M2	M3
Fixed effect					
Intercept	β_0	21.78 (0.91)	21.72 (0.84)	21.67 (0.79)	21.68 (0.80)
Treatment	β_1	−0.29 (1.29)	−0.16 (1.29)	−0.08 (1.29)	−0.09 (1.31)
Post	β_2	−4.04 (1.28)	−4.06 (1.29)	−4.03 (0.99)	−4.05 (1.01)
Interaction	β_3	−1.36 (1.81)	−1.33 (1.82)	−1.37 (1.42)	−1.35 (1.43)
Random effects					
Residual	σ^2	37.04	37.79 (4.03)	21.73 (3.32)	22.08 (4.89)
Counsellor	τ_a		−0.68 (0.44)	−1.12 (0.49)	−1.07 (0.51)
Client	τ_b			15.83 (4.52)	15.54 (4.74)
Interaction	τ_c				6.28 (7.79)

post-intervention. However, the posterior probability of a negative interaction effect $P(\beta_3 < 0 \mid \mathbf{y})$ was around 84% under M2. There was no convincing data evidence that on average the EW treatment outperformed the AW treatment.

When interpreting the estimated covariance components under M1 and M2, it can be seen that the estimated covariance among scores of clients assigned to the same counsellor was negative under the BCSM models, and around $\tau_a = -.68$ under M1. Thus, scores from clients treated by the same counsellor correlated negatively. This led to an increase of the residual variance estimate for M1 in comparison to the estimated residual variance of model M0. The residual variance was underestimated under M0, since the residuals were not independently distributed but correlated negatively. The standard deviation of the intercept was around 8% smaller under M1 in comparison to M0. The negative correlation among client scores affected the estimated standard deviation of the intercept, where the standard deviations of the other regression components under M1 were almost equal to the corresponding standard errors under M0. The dependence structure implied by the clustering of clients by counsellors cannot be represented by a counsellor random effect, since the estimated cluster correlation was negative. This makes the BCSM particularly useful to model negative cluster correlation.

When accounting for the dependence among client's pre- and post-intervention scores, the estimated covariance of τ_a was more negative under M2 than under M1 and around -1.12. This led to a further reduction of the standard deviation of the intercept to .79. This negative covariance of τ_a led to an increase of the residual variance under M2. However, the estimated positive covariance of $\tau_b = 15.83$ led to a decrease of the residual variance to 21.73. The estimated

standard deviation of the Post effect, β_2 decreased to .99 due to accounting for the correlation between client's scores. The standard deviation of the interaction effect also seriously decreased from 1.82 to 1.42. Note that the dependence structure did not influence the standard deviation of the pre- post-intervention difference between treatment groups (i.e. standard deviation of β_1).

The negative correlation among client scores from the same counsellor indicated that there was individual variability in treatment effects across the clients of the same counsellor. In the same condition and for the same counsellor, some clients benefited from the treatment, where others did not and even showed an increase in score. This phenomenon of individual treatment effects was identified by the negative cluster correlation, which was also significant when considering the 95% HPD interval under M1 and M2. The negative cluster correlation of -1.12 illustrated that there was more heterogeneity in test scores than explained by the reduction in scores at the post-intervention and the (non-significant) mean difference between the two conditions.

Post-intervention individual treatment effects

To investigate the individual treatment effect further, the model BCSM M3 was defined with a random interaction effect. This represented random variability in the treatment condition EW at the post-intervention across clients, while also accounting for the clustering by clients and counsellors. The covariance structure of M3 for the client scores of counsellor i is given by

$$\boldsymbol{\Sigma}_i = \mathbf{I}_{nb}\sigma^2 + \underbrace{diag(\mathbf{Z}_i)\tau_c}_{\text{Interaction}} + \underbrace{\mathbf{J}_{nb}\tau_a}_{\text{Counsellor}} + \underbrace{(\mathbf{I}_b \otimes \mathbf{J}_n)\,\tau_b}_{\text{Client}}, \tag{8.21}$$

and the covariance matrix is counsellor specific due to the $\mathbf{Z}_i$. However, this random interaction-effect cannot be estimated, since each client only had one observation at the post-intervention. The interaction variable $\mathbf{Z}_i$ is a diagonal matrix in the covariance matrix with τ_c a residual variance parameter. Thus, the random interaction effect implies an interaction-specific residual variance in the covariance matrix. The dependence structure in Equation (8.21) represents heteroscedastic residual variances, with σ^2 the common residual variance and $\sigma_1^2 = \sigma^2 + \tau_c$ the contribution of the random interaction variance to the common residual variance. Note that the dependence structure is extended with just one additional variance parameter representing the random interaction variance for clients in the EW condition. There is data evidence in favor of individual treatment effects of clients in the EW condition, when the residual variance σ_1^2 is greater than σ^2. In the Appendix, the posterior distribution of σ^2 and τ_c is given,

and the adjustment of the shift parameters in the posterior distributions of the other covariance parameters.

The estimates of BCSM M3 are given in Table 8.2. It can be seen that the fixed regression effects did not change when including the random interaction effect. The estimated residual variance was slightly higher. The standard deviation increased, since less observations were used to estimate the common residual variance. The estimated cluster dependence of clients and of counsellors were also around the estimated values of BCSM M2. The estimated random interaction variance was around 6.28, which showed that there was more residual variance in the post-intervention scores in the EW condition. In the Appendix it is shown that the $\tau_c \geq -\sigma^2$, and τ_c is allowed to be negative. The BCSM simply makes it possible to evaluate the data support in favor of individual variation, since the interaction variance is allowed to be negative. In this case, the interaction variance was estimated to be positive with 80% posterior probability.

The effect of the EW-treatment varied across individuals, where some benefitted more from the treatment than others. The relatively large individual variation showed that for some clients the EW-treatment was very effective but not for others. A main difference between treatments was not found partly due to this individual variation. The posterior standard deviation of the interaction variance was high and around 7.79, and around 20% of the posterior distribution of the τ_c supported negative variance values. In that case, the random interaction effect lead to a common reduction in the residual variance in the EW condition, which provide more support for a main treatment effect and less support for individual variation in the treatment effect. However, an effect of a negative variance on the standard deviation of the mean interaction effect would be very small, since this can only be accomplished through the covariance matrix of the fixed effect, where it would be absorbed by other more influential factors. Nevertheless, it can be argued that 80% posterior probability is sufficient to conclude that there is individual variation in the EW-treatment effect.

Visualizing individual treatment effects

The individual variation in treatment effect is further illustrated. In Figure 8.3, the posterior expected post-intervention scores are plotted against the expected reduction in scores for clients treated by different counsellors. It is shown that for counsellor 1 (filled squared box) and for counsellor 2 (filled circles), some clients show a reduction below the average of -4, where other clients treated by the same counsellor show an above-average reduction in scores. Clients treated by the same counsellor show a large deviation in reduced scores. This hetero-

geneity in reduced scores for clients of the same counsellor is manifested by a negative cluster correlation. This means that the level of score reductions varies across clients of the same counsellor. Therefore, it is not possible to identify a common counsellor effect, since this would imply less heterogeneity in reduced scores and a positive correlation. In fact, the effect of the counsellor varies across clients, where some clients benefitted more from the counsellor than others. This can be identified as the detection of an individualized counsellor effect. The crossed marks in Figure 8.3 represent reduced scores of the ES treatment. Although, reduced scores from clients in the ES condition of counsellor one are all below the average, some clients of counsellor two scored above average in this condition. The ES treatment is likely to be more effective for counsellor one than for counsellor two. Therefore, the individualized treatment effects of the counsellors may also include heterogeneity in the AW and EW treatments.

In Figure 8.4 (upper plot), the pre- and post-intervention scores are plotted against the fitted residuals under BCSM M2. It can be seen that the residuals are directly defined in relation to the outcome variable, and differences between residuals are caused by the effects of categorical predictor variables. This illustrates that the BCSM is a parsimonious model. Despite the complex two-way nested clustering structure, the fitted residuals can be directly explained by the differences caused by the categorical predictor variables. Under a latent variable model, the fitted residuals would have been scaled in relation to the estimated latent variables. The lower plot shows the post-intervention scores against the difference between the post- and pre-intervention residuals. The filled circles are those related to counsellor one, and the filled squares to those of counsellor two. It can be seen that for both counsellors, some clients showed a large decrease in residual value, where others did not. This heterogeneity across clients treated by the same counsellor in residual reduction from the pre-intervention to the post-intervention shows again that some clients benefited from the treatment, where others did not.

This analysis of the treatment effects was not possible with an MLM, since the factor variable counsellor implied a negative cluster correlation. This led to singular model, when using the LME4 package in R, and the random effects structure was considered too complex to be supported by the data. However, by ignoring the negative cluster correlation, relevant information was ignored. Counsellors provided individual instructions to their clients, which led to a decrease in scores for some clients but not for others. The differential treatment by counsellors was identified by the negative cluster correlation. The Bayesian estimation procedure for the BCSM did not have any issues in estimating the model parameters despite any negative clustering effects and the small sample

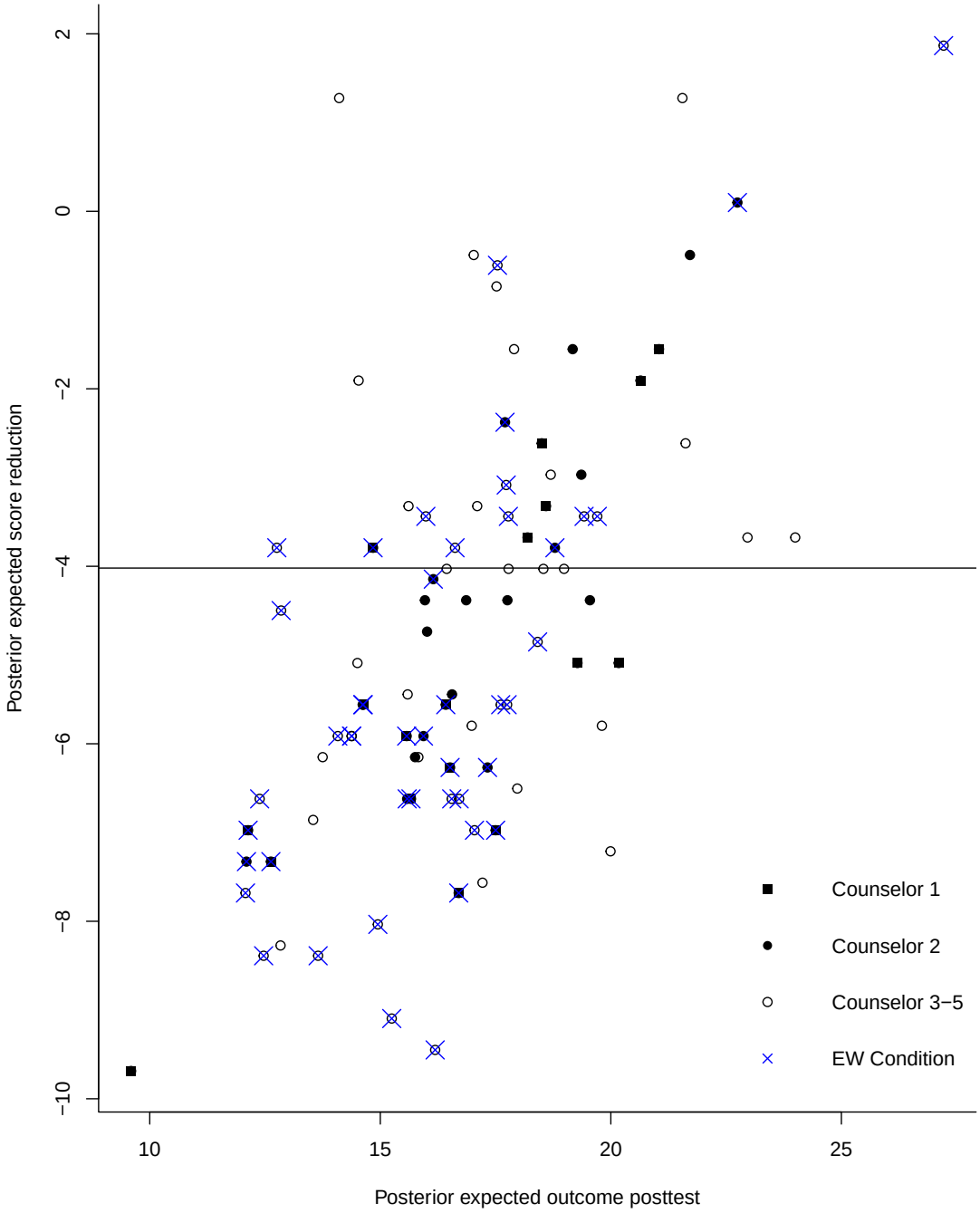


Figure 8.3: The email-counselling study from Lamers et al. (2015). Posterior expected post-intervention scores against the expected reduction in scores of clients across counsellors.

size.

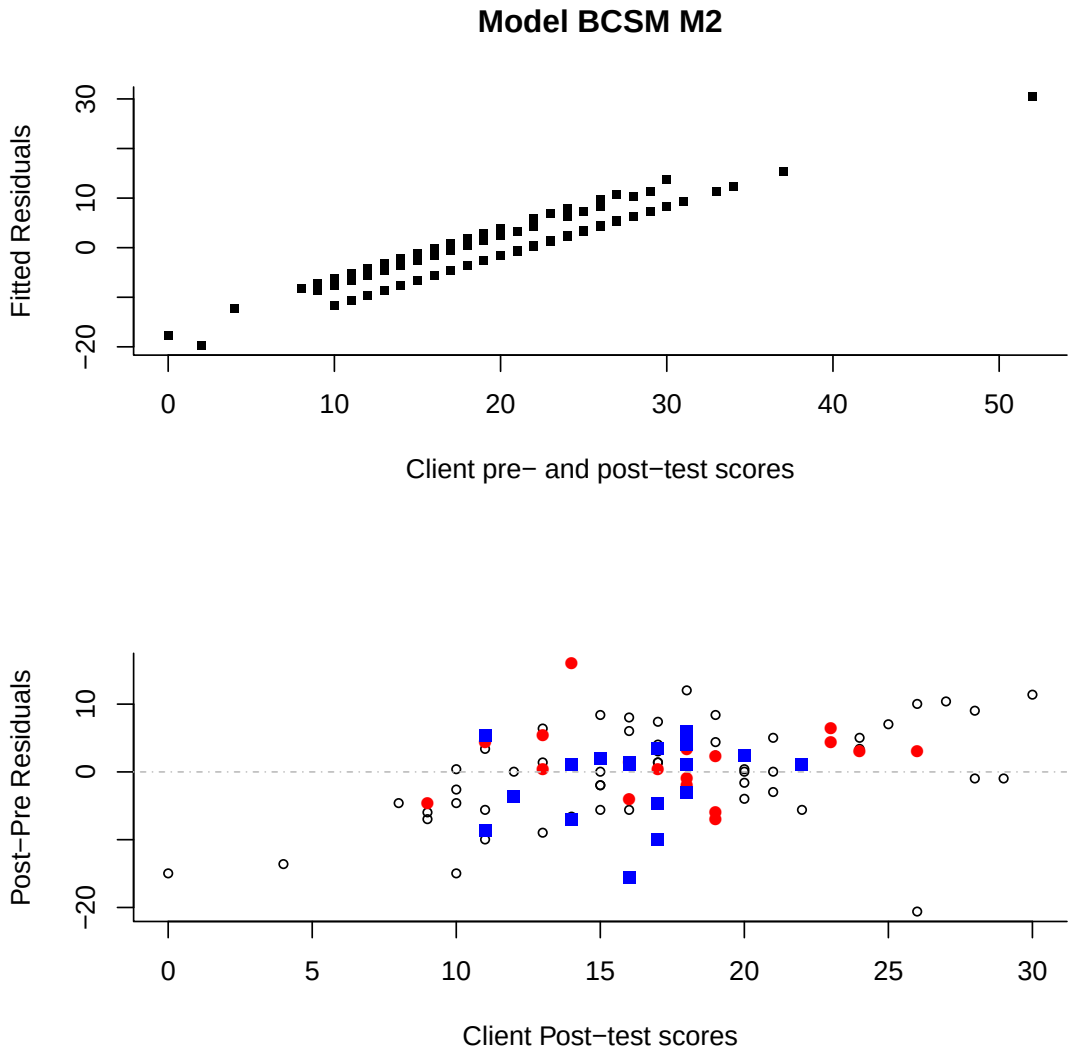


Figure 8.4: The fitted residuals under the BCSM M2, and the post-intervention scores against client's post- minus pre-intervention residuals for different counsellors.

Discussion

We introduced the novel statistical modeling framework Bayesian Covariance Structure Modelling and emphasized the understanding of BCSM, rather than discussing the underlying mathematical rigour. We designed a simulation study and analysed real data to demonstrate that BCSM can 1) assess (very) small variance components (i.e. near the lower-bound of zero), 2) assess negative variance components, 3) assess complex dependence structures given small data sets, and 4) assess individualized effects (by modelling negative associations between clustered observations). We discuss our findings, reflect on the limitations of our study, and suggest further BCSM research.

MLM software programs can produce negative variance estimates, which in general is considered to be an objectionable characteristic of the estimation methods, and limit the usefulness of variance component techniques (Thompson, 1962). For instance, the online SAS documentation (section Negative Variance Component Estimates) reports that it is common practice to treat negative variance components as if they are zero (assuming the model is appropriate for the data, see https://support.sas.com/en/documentation.html). It is argued that a larger sample size might be needed, outliers cause violations of model assumptions, or the variability is too large. However, it is also stated that negative variance estimates can indicate that clustered observations are negatively correlated. The BCSM gives support to modeling negatively correlated observations while using a very parsimonious modeling approach to make it suitable for very small data sets. From a statistical point of view, the BCSM is the natural extension of the MLM approach.

Many statistical models rely on multiple observations for proper model behaviour. Statistical modelling runs into problems when there are only a few observations (i.e. when data is sparse), yet, small samples are by no means a rare occurrence in many scientific disciplines. After all, a (relatively) small(er) data set does not imply a lesser degree of importance, as there are a variety of reasons why data sets could be small. Correct statistical modelling is arguably even more important when, for example, the population of the target group is extremely sparse (e.g., babies with a life-threatening orphan disease), difficult to access (e.g., toddlers with autism from refugees), or very costly (e.g., heart-lung transplants in infants). Small data sets are especially challenging for mixed effects models, as the sample size restrictions apply to each (modelled) hierarchical level in the data. Limited sample sizes greatly constrain meaningful statistical inference, as the sample determines the sufficient number of clusters (usually too few), and the size of the clusters themselves (usually too small). To overcome these issues, researchers often simplify their hypotheses and corresponding statistical models. Instead of doing that, our simulation study showed that BCSM can deal with few clusters with a small number of observations.

The reason why small and even negative variance components are easily estimated under BCSM, is because the so-called boundary effects can be weakened by extending the parameter space to include negative values. Usually, zero is the lower-bound of variance components because –in the standard multilevel modelling approach– a random effect is used to model dependences among treated individuals. However, the random effect variance is restricted to be positive and, as a result, always implies a positive association among individuals. Negative associations among measurements caused by the cluster (such as the counsel-

lor, or the teacher), which increases the heterogeneity among treated individuals, would require the modelling of a negative random effect variance. Under BCSM, it is straightforward to assess these effects. The covariance structure of the BCSM can represent a random effect structure, but the random effects themselves do not have to be estimated. Many individual change phenomena can be represented through a multilevel model, but these methods typically require large samples and cannot always properly model heterogeneity within clusters. An important advantage of BCSM is that the covariance structure can represent a dependence structure implied by random effects, but the effects themselves do not have to be estimated. The number of BCSM parameters is drastically lower than for the standard MLM approaches, while the interpretation does not change. Thus, BCSM allows for modelling complex theories with limited data.

Limitations

The main limitation is that data was assumed for a balanced design with a one-way or two-way random effects structure. This textbook-case is –indeed– simple, but also illustrative. We choose these (balanced) dependence structures to align with our ambition to also gently introduce the BCSM. The balanced design greatly simplifies the mathematical structure underlying our analyses. Ultimately, it is also our goal to include unbalanced designs, but to keep the scope of our current article manageable, we focused on balanced designs. Of course, because unbalanced designs are so ubiquitous in practice, the BCSM is going to be extended to unbalanced designs. Furthermore, the statistical results obtained for balanced designs will be the building blocks for unbalanced designs. Meanwhile, BCSMs have been defined for much more complex dependence structures (as can be seen in Fox et al., 2017; Klotzke & Fox, 2019a, 2019b; Mulder & Fox, 2019).

Another limitation is that we relied on the default settings of the LME4 and JAGS's estimation method. We could have also adjusted and tweaked the estimation methods for optimal performance. All things considered, we justified our choice based on the relative simplicity of the one-way random effects model. The methods should be able to perform equally well (without any adjustments) for these kind of models.

A final limitation lies within the computational efforts that are needed to estimate BCSM parameters. The Gibbs sampling procedure simply requires more computation time than standard maximum likelihood methods. Ultimately, we feel that the fact that the BCSM can estimate negative cluster correlations for relative small samples far outweighs the computational cost. Also, while this

generally true for all analysis of data: data collection (usually) takes way more time, outweighing the computational time (usually) by a large margin.

Future research

The future of research into BCSM appears to be very relevant for various long-standing statistical modeling problems. One of the foremost, is model selection. As we have shown, under BCSM, zero is 'just' another value in the parameter space of the (co)variance parameter instead of (an absolute) lower bound. In the (standard) MLM, inferences about random effect variance parameters are problematic. For instance, a random-effect variance of zero, or a negative variance estimate, can be of specific interest, but is now non-testable as both these values lie outside the boundary of the parameter space. Central to psychological research is that theories or hypotheses are often expressed in the form of several competing models (Klugkist et al., 2010; Wagenmakers & Farrell, 2004). It is also often complicated to compare models that have small variance components, as these variances lie near the lower bound, and testing near (or on) the lower-bound is known to be problematic. With the BCSM, these so-called boundary effects can be avoided, or at least weakened, by extending the parameter space to include negative values, allowing not only for a more direct, but also testable model comparison. In Bayesian hypothesis testing, hypotheses are restricted to the parameter space of the prior(s). Thus, a major improvement of the BCSM is the simple solution to have a prior distribution which gives positive support to negative and positive intra-cluster correlations to make an objective decision about the nature of the clustering.

Another interesting line of future research into BCSM is an extension to make statistical inferences from (very) small data samples: BCSM has minimal sample size requirements, since it only requires two observations to estimate the intra-cluster correlation, which is –indeed– the bare minimum of observations required to compute a variance component. Data sets in the social and medical sciences often remind us that not all data is 'Big Data': small samples are by no means a rare occurrence. A small data set does not imply a lesser degree of importance, as there are a variety of reasons why data sets could be small. Correct statistical modelling is perhaps even more important when, for example, the population of the target group is extremely sparse. Even for small data sets, researchers in the social and medical sciences often have comprehensive theories available, which lead into the direction of testing many parameters with multiple and complex dependencies. Fortunately, the complexity of the BCSM is easily controlled, since each random effect structure is modelled in a separate

layer of an additive covariance structure. Doing so is much more difficult in the MLM approach, where each random effect introduces many model parameters and the exact number of parameters depends on the fit of the model.

The final suggestion for future research concerns the estimation of individual treatment effects. It is shown that the BCSM can detect individualized treatments through negative intra-individual correlations, a next step is the estimation of the effects. Estimated BCSM residuals contain the individual-specific regression (random effect) parameters and a post-hoc estimation method is needed to estimate those random effects. For positively correlated clustered observations, these estimated effects should resemble the random effect estimates under the MLM. For negatively correlated observations, a different method is needed to estimate the individual-specific contribution.

Conclusion

We hope that we have been able to show how our BCSM approach contribute to standard multilevel modelling approaches and can be applied to evaluate individualized interventions in psychology. Even though –as Pryseley et al. (2011) pointed out– negative variance components received attention for more then half a century (starting with Chernoff, 1954; Nelder, 1954), BCSM is a new way to model directly dependences between measurements and individuals.

We strongly feel that BCSM affords the possibility of estimating rich and realistic models for psychotherapy data. Given the relative importance of this question in the psychology science, we hope that the BCSM accelerates relates research into the question of how individuals change. We hope that the BCSM that we suggested serve as a starting point for empirical analyses of individual change process research, ultimately to the benefit of not only psychological science, but especially to those that rely on the benefits of (psycho)therapy.

Appendix

The random treatment effect for the client, denoted as β_{3ij}, does not define a group effect, since each client ij has only one post-intervention observation. Thus, the design matrix for the random treatment effect, $\mathbf{Z}_i$, is a diagonal matrix, with a one for each client in the EW condition at the post-intervention and a zero otherwise. The (conditional) MLM for counsellor i can be presented as

$$\begin{aligned}
\mathbf{y}_i &= \beta_0 + \beta_1 \mathbf{Treatment}_i + \beta_2 \mathbf{Post}_i + \boldsymbol{\beta}_{3i}\mathbf{Z}_i + \boldsymbol{\beta}_{(i)}\left(\mathbf{I}_b \otimes \mathbf{1}_n\right) + \alpha_i + \mathbf{e}_i, \\
\alpha_i &\sim N(0, \tau_a) \text{ (Counsellor)} \\
\boldsymbol{\beta}_{(i)} &\sim N(0, \mathbf{I}_b\tau_b) \text{ (Clients)} \\
\boldsymbol{\beta}_{3i} &\sim N(\beta_3, \mathbf{I}_{nb}\tau_c) \text{ (Interaction)} \\
\mathbf{e}_i &\sim N(0, \mathbf{I}_{nb}\sigma^2).
\end{aligned}$$

The covariance structure implied by the random effects for the clients of counsellor i is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_i &= Var\left(\boldsymbol{\beta}_{3i}\mathbf{Z}_i\right) + Var\left(\alpha_i \mathbf{1}_{nb}\right) + Var\left(\boldsymbol{\beta}_{(i)}\left(\mathbf{I}_b \otimes \mathbf{1}_n\right)\right) + Var\left(\mathbf{e}_i\right) \\
&= \mathbf{Z}_i\mathbf{Z}_i^t\tau_c + \left(\mathbf{1}_{nb}\mathbf{1}_{nb}^t\right)\tau_a + \left(\mathbf{I}_b \otimes \mathbf{1}_n\right)\left(\mathbf{I}_b \otimes \mathbf{1}_n\right)^t\tau_b + \mathbf{I}_{nb}\sigma^2 \\
&= \mathbf{I}_{nb}\sigma^2 + \mathbf{Z}_i\mathbf{Z}_i^t\tau_c + \mathbf{J}_{nb}\tau_a + \left(\mathbf{I}_b \otimes \mathbf{J}_n\right)\tau_b \\
&= \mathbf{I}_{nb}\sigma^2 + \underbrace{diag(\mathbf{Z}_i)\tau_c}_{\text{Interaction}} + \underbrace{\mathbf{J}_{nb}\tau_a}_{\text{Counsellor}} + \underbrace{\left(\mathbf{I}_b \otimes \mathbf{J}_n\right)\tau_b}_{\text{Client}}. \qquad (8.22)
\end{aligned}$$

The posterior distribution of the residual variance σ^2 and random effect variance τ_c can be derived (Step 1). Consider the expected value of the sum of squares for the scores of clients in the AW condition,

$$\begin{aligned}
E\left(SS_{E_{AW}}\right) &= E\left(\sum_{i=1}^{a}\sum_{j\in AW}\sum_{k=1}^{n}\left(y_{ijk} - \overline{y}_{ij.}\right)^2\right) \\
&= E\left(\sum_{i=1}^{a}\sum_{j\in AW}\sum_{k=1}^{n}\left(e_{ijk} - \overline{e}_{ij.}\right)^2\right) \\
&= n_0\left(n-1\right)\sigma^2,
\end{aligned}$$

where n_0 is the number of clients in the AW condition across all counsellors. Subsequently, variance parameter σ^2 has an inverse-gamma distribution with shape parameter $(g1 + n_0(n-1))/2$ and scale parameter $(g2 + SS_{E_{AW}})/2$. The posterior distribution of the variance parameter τ_c is based on the sum of squares of the post-intervention scores of the clients in the EW condition. The expected

value is given by

$$\begin{aligned} E\left(SS_{E_{EW}}\right) &= E\left(\sum_{i=1}^{a}\sum_{j\in EW}\left(y_{ij2}-\overline{y}_2\right)^2\right) \\ &= E\left(\sum_{i=1}^{a}\sum_{j\in EW}\left(e_{ij2}-\overline{e}_2\right)^2\right) \\ &= \left(n_1-1\right)\left(\sigma^2+\tau_c\right), \end{aligned}$$

where n_1 is the number of post-intervention scores of clients in the EW condition, and $\overline{y}_2$ the average post-intervention score of all clients in the EW condition. The prior for σ^2 is an inverse-gamma with parameters g_1 and g_2. The prior for the τ_c is a shifted inverse gamma distribution, with the σ^2 as the shift parameter, and shape and scale parameter g_1 and g_2, respectively,

$$p\left(\tau_c \mid \sigma^2\right) \propto \left(\tau_c+\sigma^2\right)^{-g_1-1}\exp\left(\frac{-g_2}{\tau_c+\sigma^2}\right)$$

and $\tau_c \geq -1/\sigma^2$. The posterior distribution of variance parameter τ_c is a shifted-inverse gamma distribution with shape parameter $(g1+(n_1-1))/2$ and scale parameter $(g2+SS_{E_{EW}})/2$.

The posterior distribution of τ_a and τ_b depend on the average residual variance (see Equation (8.16) and (8.18); step 2a and step 2b). With heteroscedastic error variances within a cluster i, a (pooled) average variance parameter is defined. The average residual variance can be defined using a pooled variance parameter. Consider the average residual variance,

$$E\left(\overline{e}^2_{ij.}\right) = Var\left(\frac{e_{ij1}+e_{ij2}}{2}\right) = \begin{cases} \sigma^2/2 & \text{AW condition} \\ \left(\sigma^2+\tau_c/2\right)/2 & \text{EW condition} \end{cases}$$

This expression is used to define the average residual variance in cluster i using a pooled residual variance parameter. Let n_{0i} and n_{1i} define the number of clients in the AW and EW condition for counsellor i, respectively. It follows that,

$$\begin{aligned} E\left(\overline{e}^2_{i..}\right) &= \frac{n_{0i}\sigma^2/2+n_{1i}\left(\sigma^2+\tau_c/2\right)/2}{\left(n_{0i}+n_{1i}\right)^2} \\ &= \frac{\left(\frac{n_{0i}}{n_{0i}+n_{1i}}\right)\sigma^2+\left(\frac{n_{1i}}{n_{0i}+n_{1i}}\right)\left(\sigma^2+\tau_c/2\right)}{2\left(n_{0i}+n_{1i}\right)} \\ &= \frac{\tilde{\sigma}^2}{2\left(n_{0i}+n_{1i}\right)} = \frac{\tilde{\sigma}^2}{nb}, \end{aligned}$$

with $n=2$, and $b=n_{0i}+n_{1i}$, and $\tilde{\sigma}^2$ the pooled residual variance parameter

for cluster i. Finally, with n_0 and n_1 the total number of clients in the AW and EW condition, respectively, a general pooled residual variance parameter is defined using the weights $n_0/(n_0 + n_1)$ and the $n_1/(n_0 + n_1)$. This pooled variance parameter is used to define the shift parameter in the shifted-inverse gamma distribution of τ_a and τ_b (Step 3a and Step 3b).


What
Works
When
for
Whom

What
Works
When
for
Whom

A general discussion of the thesis

Woody Buzz, you're flying!

Buzz Lightyear This isn't flying, this is falling with style!

From the movie *Toy Story* by John Lasseter (1995)
produced at Pixar Animation Studios

9 General discussion

So we beat on, boats against
the current, borne back
ceaselessly into the past.

The Great Gatsby (1925, p. 261)
by F. Scott Fitzgerald

Introduction

We aim to advance ***T**herapeutic **C**hange **P**rocess **R**esearch* (TCPR), a field dedicated to answer the ***W**hat **W**orks **W**hen for **W**hom* (WWWW; Norcross & Wampold, 2011) question: what treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances (Paul, 1967, p. 111). In the next section we pose seven propositions that (we feel) can be substantiated based on the thesis (see Table 9.1). We use these propositions to reflect on the shortcomings and limitations of our work, and conclude with suggestions for future research.

The seven propositions of this thesis

Effects studies can establish that an effect occurred, but knowing that 'something' happened is not an answer to the WWWW question (and how change results from therapy remains a black box; Elliott, 2010, 2012). Ever since Sigmund Freud introduced the 'talking cure', conversation is the cornerstone to most forms of psychotherapy (Garfield, 2006). Given the central position of the therapeutic exchange (Imel et al., 2015), we argued that conversation is important for understanding what happens in therapy, as language gives the unique opportunity to express and communicate emotions, thoughts, motivations, and intentions (Tausczik & Pennebaker, 2010).

Because *text* is a 'data-format' of language that is straightforward to analyse, it is no surprise that TCPR has a long-standing tradition in the analysis of *tran-*

Table 9.1: Propositions of the thesis.

#	*Proposition*
1.	Different research disciplines rely on different methods for *Therapeutic Change Process Research* (TCPR).
2.	E-mails are a valuable data-source for studying the therapeutic exchange.
3.	There are different ways to model e-mail data, matching different TCPR preferences.
4.	Multilevel models are exceptionally suitable for TCPR, because of the way they can model e-mail data.
5.	Multilevel models cannot assess negative clustering effects.
6.	Negative clustering effects are the key to understanding *What Works When for Whom* (WWWW), TCPR's main research question.
7.	An approach to WWWW is *Bayesian Covariance Structure Modelling*.

scribed language. The systematic review in chapter 2 shows that the *Innovative Moments Coding Scheme* (Gonçalves et al., 2010; Gonçalves et al., 2011), the *Narrative Process Coding Scheme* (Angus et al., 1996), the *Assimilation of Problematic Experiences Scale* (Stiles et al., 1990; Stiles et al., 1991), and *Conversation Analysis* (Peräkylä, 2012; Voutilainen et al., 2011) are the often used qualitative TCPR methods to assess the 'therapeutic talk'.

The state-of-the-art review in chapter 3 identifies four research streams of (automated) text mining methods that are used specifically for TCPR (proposition 1 in Table 9.1): *change analysts* (stream A), *engineers* (B), *explorers* (C), and the *digitals* (D). In chapter 4, we show that the streams differ in research objectives, and we characterize this as a distinction between *automation* and *explication*. We use chapter 4 to show that the largest differences between these two concepts are between the *developers* of (new) text mining applications (as discussed in chapter 3), and the *users* of these text mining solutions (mainly discussed in chapter 2).

As online counselling mainly relies on e-mail as the primary means of communication (Chester & Glass, 2006; Rochlen et al., 2004), these interventions directly produce the textual interaction between client and counsellor. As the active ingredients of therapy are included in the exchange of several e-mails, we argued in chapter 5 that these e-mails contain a wealth of information about

the WWWW question (proposition 2). In this light, it is not surprising to see TCPR move towards web-based interventions. Nowadays, there is plenty of room to do so, as traditional forms of psychotherapy are increasingly complemented by online interactions between client and counsellor (Barak et al., 2008). For some web-based interventions, the client and counsellor exclusively meet online through the exchange of e-mails. There are different approaches to model e-mail data from these interventions (proposition 3): in chapter 5 we discuss models that focus on *accuracy*; in chapter 6 we discuss an *explainable* model.

In chapter 5 we show how accurate models can be used to predict drop-out in the web-based intervention of Postel (2011), which is aimed at reducing (problematic) drinking behaviour. Even though the *Linguistic Inquiry and Word Count* program (LIWC; Pennebaker, Boyd, et al., 2015) is one of the most popular text mining methods (as discussed in chapter 3), drop-out turns out to be a multidimensional construct that is complex to associate with the e-mail texts through LIWC alone (Stark, 1992). Nevertheless, by taking this approach, chapter 5 offers insight in the possibilities of working with e-mail data through accurate models and presents some preliminary findings (which stress the importance of a good working alliance between client and counsellor, distinguish between formal and informal language, and highlight the importance of Tactus' internet forum).

We detail a specific *explainable* model in chapter 6. As many phenomena of individual change can be represented through a two-level hierarchical model, *multilevel models* arise naturally for TCPR (Raudenbush & Bryk, 2002a). The first level of the model represents each clients' individual growth trajectory through the repeated measures for each client. The second level represents variables that are not repeatedly measured (or that change throughout the intervention), such as gender, income, or social economic status (Snijders & Bosker, 2012). Considering that the client and counsellor are the two pre-eminent levels of clustering (with several clients treated by a smaller number of counsellors), it becomes apparent that the psychotherapeutic practice is –essentially– a multilevelled procedure (proposition 4; Adelson & Owen, 2012; Baldwin et al., 2005; Crits-Christoph et al., 2013; Kenny & Hoyt, 2009; Marcus et al., 2009; Nissen-Lie et al., 2010).

In chapter 6 we let the first level of the multilevel model consist out of the repeated measures (i.e. a pre- and post-therapeutic score) of well-being and depression in clients who participated in the e-mail intervention of Lamers et al. (2015). The second level consists out of the clients themselves, with several traits of their writing style. We also tried to model the clients as nested within their counsellors, so that we could assess the relative counsellor effectiveness. The counsellors would then act as the third level of the model, but we found that we were unable to do so, as it turns out that it is impossible for (standard) multilevel

models to assess the negative associations between observations within clusters. Multilevel models can only assess *positive clustering effects* (positively correlated observations in clusters), and not the effects of *negative* clustering (negatively correlated observations in clusters, proposition 5).

Negative clustering effects are relevant for TCPR, as they describe the heterogeneity of a treatment, or the heterogeneous effect a counsellor has on clients. Whereas positive clustering effects describe the *similarity* among observations in clusters, negative clustering effects describe the *dissimilarity* in clusters. Dissimilarity indicates that the heterogeneity within clusters is larger than between clusters, in other words, negative clustering effects describe the divergency in clusters. In a repeated measures design, dissimilarity between observations means that each individual in a cluster has his (or her) own change trajectory. A proper assessment of this heterogeneity unveils the contribution of the cluster to the divergency. In other words, when modelling the repeated measures of clients nested within counsellors, it becomes possible to learn something about WWWW (proposition 6). Although negative clustering effects received some attention (El Leithy et al., 2016; Kenny et al., 2002; Molenberghs & Verbeke, 2007, 2011; Nelder, 1954; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003), they appear to be largely unknown (or are neglected by) the sheer majority of the research community.

In chapter 7 we show that standard multilevel models incorrectly assume that observations are independently distributed under negative clustering, and show that ignoring these effects lead to invalid standard errors and confidence intervals, and deflated type-I errors. We also discuss why negative clustering effects are impossible to model through standard multilevel models: under the multilevel modelling framework, dependencies among observations are modelled through a random effect, with a variance that determines the strength of the positive correlations among clustered observations. Modelling negative correlations between observations (i.e. the dissimilarity) would then require a random effect with a negative variance. This cannot be done with the standard applications of multilevel modelling, which leaves an important question open: how should these negative clustering effects then be assessed?

In chapter 8, we introduce an alternative framework that is capable of assessing negative clustering effects: ***B**ayesian **C**ovariance **S**tructure **M**odelling* (BCSM). As the framework can address negative clustering, we argue that the use of BCSM can show for whom the personalized treatment worked (and for whom it did not; proposition 7). A highly negative effect indicates that some clients benefit highly from the personalized treatment, where for others positive treatment effects are more difficult to realize. These negative correlations also provide in-

formation about the counsellors who are able to help clients improve. In chapter 8 we also show that we were unable to model the counsellor effectiveness in chapter 6, because the counsellor had a negative clustering effect on the clients.

Strengths and limitations

We discuss the shortcomings and strengths of our approach by reflecting on the propositions, but first we address the two assumptions that determined the orientation of our work. Our first assumption is that *the therapeutic exchange between client and counsellor lies at the heart of TCPR*. By taking this specific position, we limited our focus to forms of therapy where the therapeutic conversation between client and counsellor plays a central role. There are however also other forms of therapy. We do not reflect on the therapeutic (change) processes from *psychiatric medication*, *eye movement desensitization* and *reprocessing* (EMDR), or the beneficial effects of –for example– *physical exercise*. Given that we embedded our work in psychotherapy research (which often focuses on the elements of therapy that are under direct influence of the counsellor), we feel that this is a justifiable limitation of our scope, and one that is in line with the research goals of the field.

An alternative approach to TCPR –that would have been within the scope of our work (but was not explored in this thesis)– could focus on a specific type of therapy, such as *psychodynamic* (psychoanalytic) *psychotherapy*, *cognitive behavioural therapy*, *cognitive analytical therapy*, *interpersonal psychotherapy*, *humanistic therapy*, *family* or *couple* (systemic) *therapy*. A discussion of TCPR within the context of a specific type of therapy comes with the advantage that the results are directly embedded in the clinical practice. A more general approach could focus on the *common* effects that underlie the many different types of treatments. Mental health professionals already noted that many forms of therapy share common elements, and there is a growing body of researchers that advocates this type of TCPR (Grencavage & Norcross, 1990; Messer & Wampold, 2002). Focusing on the shared aspects of a 'general therapeutic model' has the potential to uncover the roots of therapy, and also gives insight into the WWWW question.

It was our choice to adopt a *methodological* point of view: we wanted to carefully assess the common TCPR methods and models, so we did not address specific types of therapy or variables. We are aware of the fact that we do not present the first body of work that is directed at TCPR, but we are among the firsts to advocate an interdisciplinary approach, and towards that end we discuss various qualitative, automated, and statistical approaches. With the respect

to the latter, we present a novel modelling framework that is particularly suitable for TCPR. So, even though we limited our scope we are confident that we present a coherent and interdisciplinary methodological approach to TCPR.

The second assumption that underlies our work is that *transcripts are a valuable data-source for studying the therapeutic exchange*, but –of course– therapeutic texts do not (and cannot) include all relevant behaviour. There are many different transcription standards: some require a direction transcription of what is said, for others paraphrasing is sufficient, and some standards require that all utterances (i.e. '*oh*', '*err*', '*hmm*') and changes in the tone of voice are made explicit. These different transcription standards allow for different depths of processing, but we did not reflect on these differences. Aside from that, transcripts do not include non-verbal communication such as facial expression, posture, gestures, eye contact, and touch. Any experienced counsellor could tell how important these aspects of therapy are. So, our approach to TCPR does not consider *how* something is said, instead we focused on exactly *what* is said in therapy. We preferred an in-depth study of the active ingredients in the therapeutic language over other aspects.

We will now reflect on our propositions: our first is that that *different research disciplines rely on different TCPR methods*. In chapter 2 and 3 we took the position that diversification at expense of unification is not the way forward. We re-iterate our plea for a more integrated and unified approach to TCPR: given the central role that TCPR could play in psychotherapy research, we are positive that all disciplines would benefit from a more integrated approach. With that in mind, we like to draw attention Figure 9.1.

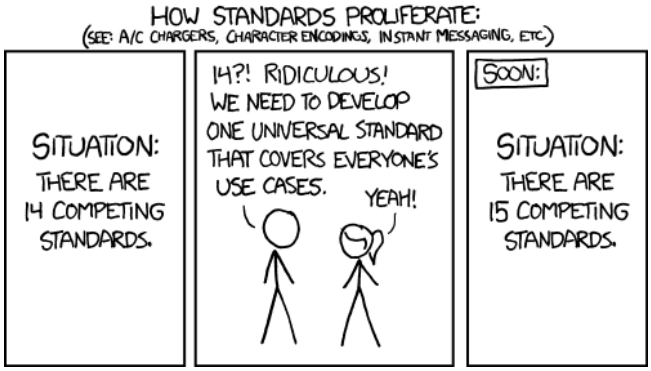


Figure 9.1: *Standards*, reprinted from xkcd comics.

Indeed, proposing to 'do something differently' (similar to proposing a new standard in Figure 9.1) comes with the risk of ending up as one of the other approaches (that set out with the same ambition). Aside from that, posing that there should be more unification is –to an extent– naive to the fact that there is

perhaps a good reason why there are so many different approaches, and maybe this differentiation is a desirable trait. As many fields rely on the analysis of language –such as for example the humanities, linguistics, social and computer sciences– it should not come as a surprise that there are equally many different approaches to do so.

We are aware of the possibility that our proposition could end up as one of the many approaches to TCPR (similar to Figure 9.1), but we do not have the impression that the TCPR methods we discussed in chapter 2 and 3 were developed with knowledge of one another (there are relatively few cross-citations and almost all approaches appear to be rooted in different areas of psychotherapeutic research). We are therefore confident that the two reviews that we presented in these chapters are a valuable addition with respect to unification of the literature (and will not contribute to more diversification). Especially because we present the *automation-explication* framework in chapter 4: in part I of the thesis we not only show that qualitative and automated approaches differ, we also show how these fields differ.

Our first proposition in part II is that *e-mails are (also) a value data-source for studying the therapeutic exchange*. Indeed, in addition to *transcriptions*, we argue in in chapter 5 and 6 that e-mails also have many of the desirable properties that *transcribed* language. However, working with e-mails also (implicitly) means that some level of technical expertise is present in the research group, as text is a difficult data-format. In addition to the difficulties of transcribed language, web-based counselling sometimes happens in the absence of any physical presence. Counselling where the client and counsellor exclusively meet online limit the counsellor in seeing and reacting to the non-verbal or voice cues of clients. Written words sometimes have a different connotation than the spoken word, and e-mails can be perceived as cold, distant and harsher than intended as a result of writing style. Therapy through written text alone has some downsides, the main of which is that the working alliance between client and counsellor is established through different mechanisms (which chapter 5 also seems to indicate).

This however is also the greatest strengths of e-mail data. As we know that web-based counselling can be effective –and client and counsellor have exclusively met via e-mail– the e-mails should contain the active ingredients of therapy. A careful assessment of these texts should give some insight in how these active ingredients bring about change over the course of therapy. As a result, the e-mail interaction could provide an important avenue for the WWWW question.

In part II we again address the *differences between research disciplines*: in part I we focused on different TCPR methods, in part II we discuss that *as the TCPR*

preferences differer between fields (mainly with respect to explication), there are different ways to model e-mail data. However, when looking at the results that we present in both chapters, we do not find large effects. In chapter 5 we study drop-out in clients with AUD, and in chapter 6 we study improvement in clients with mild depressive symptoms, and for both purposes we use the LIWC program. We chose for this program, as the majority of the specific TCPR instruments is only available for the English language (and not for Dutch, which is the main language of the datasets we used in chapter 5 and 6), others demonstrated the validity of LIWC in Dutch (Boot et al., 2017; van Wissen & Boot, 2017), and LIWC encompasses many aspects of the e-mail texts (such as function words and punctuation). Perhaps we did not find large effects because the phenomena that we studied are quite nuanced and context-dependent, and LIWC was developed as a general tool aimed at general use-cases (for example, LIWC does not contain dictionaries that are related to AUD).

Another shortcoming that underlies part II is that we do not return to the qualitative methods that we mentioned in part I, as we posed that *multilevel models are particularly useful for TCPR*. Our main reason for doing so was that we spent our time developing the framework that we presented in part III. This choice came with the downside that we could not further develop qualitative research models.

Chapter 7 demonstrates that *multilevel models cannot assess negative clustering effects*, however it is difficult to make general claims because negative clustering is not addressed uniformly in the literature. We present a simulation study that assesses the effect under various circumstances, but –to the best of our knowledge– related concepts –such as the negativity of certain variance components, or the modelling of covariance matrices under the *Structural Equation Modelling* framework– received some attention, but we are confident that overall negative clustering effects are largely unknown to the sheer majority of the research community.

The few who wrote about the topic (El Leithy et al., 2016; Kenny et al., 2002; Molenberghs & Verbeke, 2007, 2011; Nelder, 1954; Oliveira et al., 2017; Pryseley et al., 2011; Verbeke & Molenberghs, 2003), do so with great diversity (with few attempts of bringing these closer together). We feel that the overview in chapter 7 provides a useful basis for a large and comprehensive investigation into negative clustering effects. We hope that others find the lessons that we present at the end of chapter 7 to be helpful.

In chapter 8 we claim that *negative clustering effects are the key to understanding for WWWW, the main research question TCPR*, but this is a 'very quantitative' solution, with no widely available software. In part I, we discussed qualitative TCPR

methods, and in part III we propose a quantitative model. Although we adopted an interdisciplinary approach, we are aware that because of its technical nature the BCSM model is –in its current form– not accessible to all.

We have to make another reservation: even though we present BCSM as a solution, we do not present software that can be used by researchers outside of our team. So, in its current form, it is not possible for others to also apply BCSM to their own work. Eventually, all ambitions in research projects end up facing some time constraints, BCSM would have benefited from having software available so that others can apply BCSM to they own models. This brings us to our last limitation: we presented *BCSM as an approach to WWWW*, but as of right now BCSM only works for balanced datasets (which means that BCSM is restricted to datasets where there are an equal number of observations for each group). For BCSM to really demonstrate its capacities, ultimately BCSM also has to address unbalanced designs.

Even though we were aware of these limitations, we had two reasons for presenting chapter 8 the way we did. First, we had to keep the scope of chapter 8 manageable, and we focused on a basic model in a balanced setting. Second, the mathematical derivations for balanced designs are less complex, which is why we addressed balanced designs first. Although we do not provide the evidence for this claim in the thesis, BCSM is not restricted to balanced designs in any way, but we have not published about these results as of yet (but this is of course one of our future ambitions).

Main implications and future directions

We hope that we have been able to show how BCSM contributes standard multilevel modelling approaches, and can be applied to evaluate personalized interventions for TCPR. Even though –as Pryseley et al. (2011) pointed out– negative clustering effects received attention for more than half a century (starting with Chernoff, 1954; Nelder, 1954), BCSM presents a novel opportunity to directly model the dependences between measurements and individuals, and has several advantages over the other approaches that are available. We strongly feel that BCSM allows for the possibility of estimating rich and realistic models for psychotherapy data. However, our purpose was –for now– not to examine these personalized treatment effects per se, but rather demonstrate that BCSM is able to model such effects. An examination into the effect of each individual would require more experimental data (which requires software that is not yet developed). Given the relative importance of the WWWW question in the psychology

science, we hope that BCSM accelerates research into the question of how individuals change. BCSM can serve as a starting point for empirical analyses of personalized treatment effects for TCPR, not only to the benefit of the psychological science, but to everyone –public health officials, clinicians, and of course the clients and patients– who relies on the benefits of (psycho)therapy.

In chapter 3, we identified four streams of text mining TCPR. Mainly for the *change analysts* of stream A and the *explorers* of stream C will BCSM be an outcome. These streams are plagued by two types of problems: the theoretical complexity of the constructs they wish to explore is large, and often greater than what the sample size allows for. BCSM is especially suitable for these types of problems. BCSM has minimal sample size requirements, since it only requires two observations to estimate the intra-cluster correlation, which is –indeed– the bare minimum of observations required to compute a variance component. Secondly, the complexity of the BCSM is easily controlled, since each random effect structure is modelled in a separate layer of an additive covariance structure. This property of BCSM allows for the rich and in-depth exploration of the theoretical constructs that are required for stream A and C, with less restrictions in sample size and model complexity. Doing so is much more difficult under the standard multilevel modelling approach, where each random effect introduces many model parameters and the exact number of parameters depends on the fit of the model.

The engineers of stream B and the digitals of stream D use large datasets. One of the elegant aspects of the BCSM is that the properties that make the framework so applicable for small datasets, also work favourably for large datasets. The main advantage of BCSM for these disciplines is that the parameter complexity does not increase when the amount of data grows. BCSM is even suitable for exponentially large datasets with repeated measures for more than a million respondents (which –again– would be more difficult to do under the standard multilevel modelling approach). Although we do not present empirical evidence for this claim in the thesis, it is straightforward to see why this is the case, as the computation of variance components is based on the sum of squares and data transformations that can be calculated regardless of the size of the data. This is a welcome addition for stream B and D, who work with datasets of considerable size. These TCPR researchers also often work with accurate models, because the complexity of the constructs is usually too large for models that are focused on explanation. As model complexity is less of an issue under BCSM, the framework could be an attractive option for researchers who wish to use models that are well-explainable.

Conclusion

We reflected on the various models and methods that can assess the therapeutic exchange between client and counsellor. By using the automation-explication framework, we conclude that fields differ in the TCPR preferences for modelling e-mail data. We argue that multilevel models are exceptionally suitable for TCPR, because these models produce explainable relations between the outcome of therapy and aspects of the therapeutic interaction. We posed BCSM as an alternative with several attractive model features for TCPR: BCSM is suitable for the analysis of small data samples (it is possible to model cluster effects with only two observations in each cluster), BCSM can assess complex dependencies (so that multifaceted theoretical constructs can be assessed, even with a small dataset), and –unlike the standard multilevel models– BCSM can assess negative clustering effects. Through a simulation study and by analysis of a real data example, we argued that negative clustering effects are in fact personalized effects. In doing so, we presented a theoretical foundation that allows for further examining individualized effects in pursuit of the WWWW question.

What
Works
When
for
Whom

Appendices

References

Aarts, B. (2001). Corpus linguistics, Chomsky and fuzzy tree fragments. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 5–13). Rodopi.

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: A systematic literature review. *International Journal of Methods in Psychiatric Research*, *25*(2), 86–100. https://doi.org/10.1002/mpr.1481

Adelson, J. L., & Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy*, *49*(2), 152–162. https://doi.org/10.1037/a0023990

Adler, J. M. (2012). Living into the story: Agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of Personality and Social Psychology*, *102*(2), 367–389. https://doi.org/10.1037/a0025289

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114. https://doi.org/10.2307/271083

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, *4*, 463–476. https://doi.org/10.1162/tacl_a_00111

Amichai-Hamburger, Y., Klomek, A. B., Friedman, D., Zuckerman, O., & Shani-Sherman, T. (2014). The future of online therapy. *Computers in Human Behavior*, *41*(1), 288–294. https://doi.org/10.1016/j.chb.2014.09.016

Anderson, T., Bein, E., Pinnell, B. J., & Strupp, H. H. (1999). Linguistic Analysis of Affective Speech in Psychotherapy: A case grammar approach. *Psychotherapy Research*, *9*(1), 88–99. https://doi.org/10.1080/10503309912331332611

Andersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cognitive Behaviour Therapy*, *38*(4), 196–205. https://doi.org/10.1080/16506070903318960

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman, E. (2014). Guided Internet-based vs. face-to-

face cognitive behavior therapy for psychiatric and somatic disorders: A systematic review and meta-analysis. *World Psychiatry*, *13*(3), 288–295. https://doi.org/10.1080/16506070903318960

Andrade, A. L. M., Caraveo-Anduaga, J. J., Berglund, P., Bijl, R., Kessler, R. C., Demler, O., Walters, E., Kýlýç, C., Offord, D., Üstün, T. B., & Wittchen, H. U. (2013). Cross-national comparisons of the prevalences and correlates of mental disorders. *Bulletin of the World Health Organization*, *78*(4), 413–426. https://doi.org/10.1590/S0042-96862000000400003

Andrade, A. L. M., de Lacerda, R. B., Gomide, H. P., Ronzani, T. M., Sartes, L. M. A., Martins, L. F., Bedendo, A., Souza-Formigoni, M. L. O., Vromans, I. S., Poznyak, V., Fitzmaurice, G., Rekve, D., Martin Abello, K., Kramer, J., Rosier, I., Tiburcio-Sainz, M., Lara, M. A., Padruchny, D., Ambekar, A., … Schaub, M. P. (2016). Web-based self-help intervention reduces alcohol consumption in both heavy-drinking and dependent alcohol users: A pilot study. *Addictive Behaviors*, *63*(1), 63–71. https://doi.org/10.1016/j.addbeh.2016.06.027

Andrews, G., Cuijpers, P., Craske, M. G., McEvoy, P., & Titov, N. (2010). Computer Therapy for the Anxiety and Depressive Disorders Is Effective, Acceptable and Practical Health Care: A Meta-Analysis (B. T. Baune, Ed.). *PLOS One*, *5*(10), e13196. https://doi.org/10.1371/journal.pone.0013196

Andrews, G., Issakidis, C., Sanderson, K., Corry, J., & Lapsley, H. (2004). Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment comparison of the cost-effectiveness of treatment of ten mental disorders. *The British Journal of Psychiatry*, *184*(6), 526–533. https://doi.org/10.1192/bjp.184.6.526

Angus, L. E., & Greenberg, L. S. (2011). *Working with narrative in emotion-focused therapy: Changing stories, healing lives.* American Psychological Association. https://doi.org/10.1037/12325-000

Angus, L. E., Hardtke, K., & Levitt, H. (1996). *Narrative Processes Coding System training manual*. Unpublished manuscript, Angus Narrative Lab, York University, Toronto, Ontario, Canada.

Angus, L. E., Levitt, H., & Hardtke, K. (1999). The narrative processes coding system: Research applications and implications for psychotherapy practice. *Journal of Clinical Psychology*, *55*(10), 1255–1270. https://doi.org/10.1002/(sici)1097-4679(199910)55:10<1255::aid-jclp7>3.0.co;2-f

Angus, L. E., & McLeod, J. (2004). *The Handbook of Narrative and Psychotherapy: Practice, Theory, and Research*. Sage Publications. https://doi.org/10.4135/9781412973496

Arntz, A., Hawke, L. D., Bamelis, L., Spinhoven, P., & Molendijk, M. L. (2012). Changes in natural language use as an indicator of psychotherapeutic change in personal-

ity disorders. *Behaviour Research and Therapy*, *50*(3), 191–202. https://doi.org/10.1016/j.brat.2011.12.007

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*(5), 816–827. https://doi.org/10.1037/a0029607

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 9–49. https://doi.org/10.1186/1748-5908-9-49

Back, M. D., Egloff, B., & Küfner, A. C. P. (2011). and Lessons Learned for the Analysis of Large Digital Data Sets Article in Psychological Science. *Psychological Science*, *22*(6), 837–838. https://doi.org/10.1177/0956797611409592

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151–164. https://doi.org/10.1037/a0030642

Baldwin, S. A., & Imel, Z. E. (2013). Therapist Effects. In M. J. Lambert (Ed.), *Handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). John Wiley & Sons, Ltd.

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, *73*(5), 924–935. https://doi.org/10.1037/0022-006X.73.5.924

Baldwin, S. A., Stice, E., & Rohde, P. (2008). Statistical analysis of group-administered intervention data: Reanalysis of two randomized trials. *Psychotherapy Research*, *18*(4), 365–376. https://doi.org/10.1080/10503300701796992

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the Alliance-Outcome Correlation: Exploring the Relative Importance of Therapist and Patient Variability in the Alliance. *Journal of Consulting and Clinical Psychology*, *75*(6), 842–852. https://doi.org/10.1037/0022-006X.75.6.842

Ball, S. A., Carroll, K. M., Canning-Ball, M., & Rounsaville, B. J. (2006). Reasons for dropout from drug abuse treatment: Symptoms, personality, and motivation. *Addictive Behaviors*, *31*(2), 320–330. https://doi.org/10.1016/j.addbeh.2005.05.013

Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, 26–33. https://doi.org/10.3115/1073012.1073017

Barak, A., Hen, L., Boniel-Nissim, M., & Shapira, N. (2008). A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. *Journal of Technology in Human Ser-*

vices, *26*(2-4), 109–160. https://doi.org/10.1080/15228830802094429

Barcikowski, R. S. (1981). Statistical Power with Group Mean as the Unit of Analysis. *Journal of Educational Statistics*, *6*(3), 267–285. https://doi.org/10.2307/1164877

Barkham, M., Stiles, W. B., & Shapiro, D. A. (1993). The shape of change in psychotherapy: longitudinal assessment of personal problems. *Journal of Consulting and Clinical Psychology*, *61*(4), 667–677. https://doi.org/10.1037/0022-006X.61.4.667

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational Research*, *45*(2), 143–154. https://doi.org/10.1080/0013188032000133548

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(22), 9977–9982. https://doi.org/10.1073/pnas.92.22.9977

Beekman, A. T. F., Deeg, D. J. H., van Limbeek, J., Braam, A. W., de Vries, M. Z., & van Tilburg, W. (1997). Criterion Validity of the Center for Epidemiologic Studies Depression scale (CES-D): Results from a Community-Based Sample of Older Subjects in the Netherlands. *Psychological Medicine*, *27*(1), 231–235. https://doi.org/10.1017/S0033291796003510

Bennett, K., Bennett, A. J., & Griffiths, K. M. (2010). Security considerations for e-mental health interventions. *Journal of Medical Internet Research*, *12*(5), e61. https://doi.org/10.2196/jmir.1468

Berger, M., Wagner, T. H., & Baker, L. C. (2005). Internet use and stigmatized illness. *Social Science and Medicine*, *61*(8), 1821–1827. https://doi.org/10.1016/j.socscimed.2005.03.025

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python* (1st ed.). O'Reilly Media. https://doi.org/10.2307/40925581

Birren, J. E., & Deutchman, D. E. (1991). *Guiding autobiography groups for older adults: Exploring the fabric of life*. John Hopkins University Press. https://doi.org/10.1353/book.3469

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://doi.org/10.1016/c2009-0-22409-3

Bishop, L. (2009). Ethical Sharing and Reuse of Qualitative Data. *Australian Journal of Social Issues*, *44*(3), 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Blankers, M. (2011). *E-Mental Health Interventions for Harmful Alcohol Use: Research Methods and Outcomes* (Doctoral dissertation). University of Amsterdam.

Blei, D. M., Ng, A. Y.-T., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(1), 993–1022.

Bluck, S., & Levine, L. J. (1998). Reminiscence as autobiographical memory: a catalyst for reminiscence theory development. *Ageing and Society*, *18*(2), 185–208. https://doi.org/10.1017/s0144686x98006862

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, *31*(4), 507–534. https://doi.org/10.1007/bf02289521

Bohlmeijer, E. T., & Westerhof, G. J. (2010). *Op verhaal komen. Je autobiografie als bron van wijsheid [The stories we live by]*. Uitgeverij Boom. https://doi.org/10.1007/bf03089415

Boldrini, T., Nazzaro, M. P., Genova, F., & Gazzillo, F. (2017). Mentalization as a Predictor of Psychoanalytic Outcome: An Empirical Study of Transcribed Psychoanalytic Sessions Through the Lenses of a Computerized Text Analysis Measure of Reflective Functioning. *Psychoanalytic Psychology*, *35*(2), 196–204. https://doi.org/10.1037/pap0000154

Bonab, H. R., & Can, F. (2016). A theoretical framework on the ideal number of classifiers for online ensembles in data streams. *International Conference on Information and Knowledge Management, Proceedings*, *24-28-Octo*, 2053–2056. https://doi.org/10.1145/2983323.2983907

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, *6*(1), 65–76. https://doi.org/10.1075/dujal.6.1.04boo

Braakmann, D. (2015). Historical Paths in Psychotherapy Research. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research: Foundations, process, and outcome* (1st ed., pp. 39–65). Springer. https://doi.org/10.1007/978-3-7091-1382-0_3

Breiman, L. (2001a). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213725

Brewin, C. R. (2006). Understanding cognitive behaviour therapy: A retrieval competition account. *Behaviour Research and Therapy*, *44*(6), 765–784. https://doi.org/10.1016/J.BRAT.2006.02.005

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American statistical Association*, *58*(301), 85–96. https://doi.org/10.2307/2282956

Bruffaerts, R., Bonnewyn, A., & Demyttenaere, K. (2007). Delays in seeking treatment for mental disorders in the Belgian general population. *Social Psychiatry and Psychiatric Epidemiology*, *42*(11), 937–944. https://doi.org/10.1007/s00127-007-0239-3

Brzozowski, J. A. (1964). Derivatives of Regular Expressions. *Journal of the Association for Computing Machinery*, *11*(4), 481–494. https://doi.org/10.1145/321239.321249

Bucci, W. (2013). The referential process as a common factor across treatment modalities. *Research in Psy-*

chotherapy: Psychopathology, Process and Outcome, *16*(1), 16–23. https://doi.org/10.7411/RP.2013.003

Bucci, W., & Maskit, B. (2006). A Weighted Referential Activity Dictionary. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theory and applications* (pp. 49–60). Springer Netherlands. https://doi.org/10.1007/1-4020-4102-0_6

Bucci, W., & Maskit, B. (2007). Beneath the surface of the therapeutic interaction: The psychoanalytic method in modern dress. *Journal of the American Psychoanalytic Association*, *55*(4), 1355–1397. https://doi.org/10.1177/000306510705500412

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261304. https://doi.org/10.1177/0049124104268644

Butler, R. N. (1963). The Life Review: An Interpretation of Reminiscence in the Aged. *Psychiatry*, *26*(1), 65–76. https://doi.org/10.1080/00332747.1963.11023339

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*. https://doi.org/10.1017/S1351324916000383

Campbell, L. F., Norcross, J. C., Vasquez, M. J. T., & Kaslow, N. J. (2013). Recognition of psychotherapy effectiveness: the APA resolution. *Psychotherapy*, *50*(1), 98. https://doi.org/10.1521/psyc.2009.72.1.32

Campbell, R., & Pennebaker, J. W. (2003). The secret life of pronouns: flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Can, D., Atkins, D. C., & Narayanan, S. (2015). A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 339–343.

Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. (2012). A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features. *Thirteenth Annual Conference of the International Speech Communication Association.*, 2–5.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2016). "It sounds like ...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343–350. https://doi.org/10.1037/cou0000111

Cariola, L. A. (2015). Semantic Expressions Of The Body Boundary Personality In Person Centered Psychotherapy. *International Body Pyschotherapy Journal*, *14*(1), 48–64. https://doi.org/10.1080/17432979.2010.530060#.VBfsNC6wJRU

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., Fenton, L., & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence*, *57*(3), 225–238. https://doi.org/10.1016/S0376-8716(99)00049-6

Carroll, L. (1865). *Alice's Adventures in Wonderland*. Macmillan Publishers. https://doi.org/10.2307/j.ctvc7785k

Chapman, C., Slade, T., Hunt, C., & Teesson, M. (2015). Delay to first treatment contact for alcohol use disorder. *Drug and Alcohol Dependence*, *147*, 116–121. https://doi.org/10.1016/j.drugalcdep.2014.11.029

Chen, E. E., & Wojcik, S. P. (2016). A Practical Guide to Big Data Research in Psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 573–578. https://doi.org/10.1214/aoms/1177728725

Chester, A., & Glass, C. A. (2006). Online counselling: A descriptive analysis of therapy services on the Internet. *British Journal of Guidance and Counselling*, *34*(2), 145–160. https://doi.org/10.1080/03069880600583170

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, *2*(3), 113–124.

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, *2*, 128–137. https://doi.org/10.1109/IRI.2012.6302998

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research*, *11*(2), 1–17. https://doi.org/10.2196/jmir.1194

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Frontiers of social psychology: Social communication* (1st ed., pp. 343–359). Psychology Press.

Clarke, G., Eubanks, D., Reid, E., Kelleher, C., O'Connor, E., DeBar, L. L., Lynch, F., Nunley, S., & Gullion, C. (2005). Overcoming Depression on the Internet (ODIN) (2): a randomized trial of a self-help depression skills program with reminders. *Journal of Medical Internet Research*, *7*(2), 1–12. https://doi.org/10.2196/jmir.7.2.e16

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, *62*(8), 752–758. https://doi.org/10.1136/jech.2007.060798

Cloud, R. N., & Peacock, P. L. (2001). Internet screening and interventions for problem drinking: Results from the www.carebetter.com pilot study. *Alcoholism Treatment Quar-*

terly, *19*(2), 23–44. https://doi.org/10.1300/J020v19n02_02

Copeland, J., & Hall, W. (1992). A comparison of predictors of treatment drop-out of women seeking drug and alcohol treatment in a specialist women's and two traditional mixed-sex treatment services. *Addiction*, *87*(6), 883–890. https://doi.org/10.1111/j.1360-0443.1992.tb01983.x

Copeland, J., & Martin, G. (2004). Web-based interventions for substance use disorders. *Journal of Substance Abuse Treatment*, *26*(2), 109–116. https://doi.org/10.1016/S0740-5472(03)00165-X

Cremers, A., & Ginsburg, S. (1975). Context-free grammar forms. *Journal of Computer and System Sciences*, *11*(1), 86–117. https://doi.org/10.1016/S0022-0000(75)80051-1

Crits-Christoph, P., Gibbons, M. B. C., & Mukherjee, D. (2013). Psychotherapy Process-Outcome Research. In M. J. Lambert (Ed.), *Handbook of psychotherapy and behavior change* (6th ed., pp. 298–340). John Wiley & Sons, Ltd.

Crowder, M. J., & Hand, D. J. (1990). *Analysis of Repeated Measures* (1st ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781315137421

Cunningham, J. A., & Breslin, F. C. (2004). Only one in three people with alcohol abuse or dependence ever seek treatment. *Addictive Behaviors*, *29*(1), 221–223. https://doi.org/10.1016/S0306-4603(03)00077-7

Cunningham, J. A., Sobell, L. C., Sobell, M. B., & Gaskin, J. (1994). Alcohol and drug abusers' reasons for seeking treatment. *Addictive Behaviors*, *19*(6), 691–696. https://doi.org/10.1016/0306-4603(94)90023-X

Degenhardt, L., Charlson, F., Ferrari, A., Santomauro, D., Erskine, H., Mantilla-Herrara, A., Whiteford, H., Leung, J., Naghavi, M., Griswold, M., Rehm, J., Hall, W., Sartorius, B., Scott, J., Vollset, S. E., Knudsen, A. K., Haro, J. M., Patton, G., Kopec, J., … Vos, T. (2018). The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Psychiatry*, *5*(12), 987–1012. https://doi.org/10.1016/S2215-0366(18)30337-7

de Graaf, R., ten Have, M. L., & van Dorsselaer, S. (2010). *De psychische gezondheid van de Nederlandse bevolking: NEMESIS-2: opzet en eerste resultaten*. Trimbos-instituut [host].

de Graaf, R., Tuithof, M., van Dorsselaer, S., & ten Have, M. (2011). *Verzuim door psychische en somatische aandoeningen bij werkenden Resultaten van de 'Netherlands Mental Health Survey and Incidence Study-2' (NEMESIS-2)* (tech. rep.). Trimbos-instituut. Utrecht.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Jour-*

nal of Machine Learning Research, *14*, 2349–2353.

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*. https://doi.org/10.1017/pan.2017.44

Donoho, D. (2017). 50 years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Dorman, J. P. (2008). The effect of clustering on statistical tests: An illustration using classroom environment data. *Educational Psychology*, *28*(5), 583–595. https://doi.org/10.1080/01443410801954201

Duval, A. (2019). Explainable Artificial Intelligence (XAI). *Unpublished thesis: The University of Warwick*. https://doi.org/10.13140/RG.2.2.24722.09929

El Leithy, H. A., Abdel Wahed, Z. A., & Abdallah, M. S. (2016). On non-negative estimation of variance components in mixed linear models. *Journal of Advanced Research*, *7*(1), 59–68. https://doi.org/10.1016/J.JARE.2015.02.001

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review*, *77*(3), 378–394. https://doi.org/10.1111/j.1751-5823.2009.00092.x

Elliott, R. (2010). Psychotherapy change process research: Realizing the promise. *Psychotherapy Research*, *20*(2), 123–135. https://doi.org/10.1080/10503300903470743

Elliott, R. (2012). Qualitative Methods for Studying Psychotherapy Change Processes. In A. R. Thompson & D. Harper (Eds.), *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners* (1st ed., pp. 69–111). Wiley-Blackwells. https://doi.org/10.1002/9781119973249.ch6

Eubanks, C. F., Burckell, L. A., & Goldfried, M. R. (2018). Clinical consensus strategies to repair ruptures in the therapeutic alliance. *Journal of Psychotherapy Integration*, *28*(1), 60–76. https://doi.org/10.1037/int0000097

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, *7*(1), 1–11. https://doi.org/10.2196/jmir.7.1.e11

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal*, *22*(2), 338–342.

Farvolden, P., Denisoff, E., Selby, P., Bagby, R. M., & Rudy, L. (2005). Usage and longitudinal effectiveness of a web-based self-help cognitive behavioral therapy program for panic disorder. *Journal of Medical Internet Research*, *7*(1). https://doi.org/10.2196/jmir.7.1.e7

Fast, L. A., & Funder, D. C. (2008). Personality as Manifest in Word Use: Correlations With Self-Report, Acquaintance Report, and Behavior.

Journal of Personality and Social Psychology, *94*(2), 334–346. https://doi.org/10.1037/0022-3514.94.2.334

Feldman, J., & Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain and language*, *89*(2), 385–392. https://doi.org/10.1016/S0093-934X(03)00355-9

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. https://doi.org/10.1179/1465312512Z.00000000017

Fernández-Álvarez, J., Díaz-García, A., González-Robles, A., Baños, R., García-Palacios, A., & Botella, C. (2017). Dropping out of a transdiagnostic online intervention: A qualitative analysis of client's experiences. *Internet Interventions*, *10*, 29–38. https://doi.org/10.1016/j.invent.2017.09.001

Fitzgerald, F. S. K. (1925). *The Great Gatsby*. Charles Scribner's Sons.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). John Wiley & Sons. https://doi.org/10.1080/10543406.2013.789817

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D. C., Creed, T. A., & Narayanan, S. (2018). Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions. *Interspeech*, (September), 1908–1912. https://doi.org/10.21437/Interspeech.2018-1518

Fontao, M. I., & Mergenthaler, E. (2008). Therapeutic factors and language patterns in group therapy application of computer-assisted text analysis to the examination of microprocesses in group therapy: Preliminary findings. *Psychotherapy Research*, *18*(3), 345–354. https://doi.org/10.1080/10503300701576352

Formánek, T., Kagström, A., Cermakova, P., Csémy, L., Mladá, K., & Winkler, P. (2019). Prevalence of mental disorders and associated disability: Results from the cross-sectional CZEch mental health Study (CZEMS). *European Psychiatry: the Journal of the Association of European Psychiatrists*, *60*, 1–6. https://doi.org/10.1016/j.eurpsy.2019.05.001

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes Factor Covariance Testing in Item Response Models. *Psychometrika*, *82*(4), 979–1006. https://doi.org/10.1007/s11336-017-9577-6

Franke, B., Plante, J. F., Roscher, R., Lee, E. s. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., & Reid, N. (2016). Statistical Inference, Learning and Models in Big Data. *International Statistical Review*, *84*(3), 371–389. https://doi.org/10.1111/insr.12176

Gainsbury, S., & Blaszczynski, A. (2011). A systematic review of Internet-based therapy for the treatment of addictions. *Clinical Psychology Review*, *31*(3), 490–498. https://doi.org/10.1016/j.cpr.2010.11.007

Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A Study of Clustered Data and Approaches to Its Analysis. *Journal of Neuroscience*, *30*(32), 10601–10608. https://doi.org/10.1523/JNEUROSCI.0362-10.2010

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., Cruden, G., & Brown, C. H. (2015). Blending Qualitative and Computational Linguistics Methods for Fidelity Assessment: Experience with the Familias Unidas Preventive Intervention. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 574–585. https://doi.org/10.1007/s10488-014-0538-4

Garfield, S. L. (2006). *Therapies - modern and popular: Psyc CRITIQUES 2006*. American Medical Association.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/b16018

Gelo, O. C. G., & Manzo, S. (2015). Quantitative approaches to treatment process, change process, and process-outcome research. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research* (pp. 247–277). Springer. https://doi.org/10.1007/978-3-7091-1382-0_13

Gelo, O. C. G., Rieken, B., & Pritz, A. (2015). *Psychotherapy research: Foundations, process, and outcome*. Springer. https://doi.org/10.1007/978-3-7091-1382-0

Gelo, O. C. G., Salcuni, S., & Colli, A. (2012). Text analysis within quantitative and qualitative psychotherapy process research: Introduction to special issue. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *15*(2), 45–53. https://doi.org/10.7411/RP.2012.005

Giberson, T. R., Resick, C. J., & Dickson, M. W. (2005). Embedding leader characteristics: an examination of homogeneity of personality and values in organizations. *Journal of Applied Psychology*, *90*(5), 1002. https://doi.org/10.1037/0021-9010.90.5.1002

Gibson, J., Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. (2017). Attention networks for modeling behaviors in addiction counseling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.21437/Interspeech.2017-218

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. (2016). A deep learning approach to modeling empathy in addiction counseling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.21437/Interspeech.2016-554

Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., & Narayanan, S. (2015). Predicting Therapist Empathy in Motivational Interviews using Language Features Inspired by Psycholinguistic Norms. *Sixteenth Annual Conference of the International Speech Communication Association.*

Giesler, L. (2019). Referential Activity in the E-Therapy Program "Look at Your Drinking" – A Text-Mining Approach. *Unpublished bachelorthesis: University of Twente*.

Ginn, S., & Horder, J. (2012). "One in four" with a mental health problem: the anatomy of a statistic. *The BMJ*, *344*, e1302. https://doi.org/10.1136/bmj.e1302

Golder, S. A., & Macy, M. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, *40*, 129–152. https://doi.org/10.1146/annurev-soc-071913-043145

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). Wiley. https://doi.org/10.1002/9780470973394

Gonçalves, M. M., Matos, M., & Santos, A. (2009). Narrative therapy and the nature of "innovative moments" in the construction of change. *Journal of Constructivist Psychology*, *22*(1), 1–23. https://doi.org/10.1080/10720530802500748

Gonçalves, M. M., Ribeiro, A. P., Matos, M., Santos, A., & Mendes, I. (2010). Innovative moments coding system: a methodological procedure for tracking changes in psychotherapy. *YIS: Yearbook of idiographic science*, *2*, 107–130. https://doi.org/10.1080/10503307.2011.560207

Gonçalves, M. M., Ribeiro, A. P., Mendes, I., Matos, M., & Santos, A. (2011). Tracking novelties in psychotherapy process research: The innovative moments coding system. *Psychotherapy Research*, *21*(5), 497–509.

Gottschalk, L. A. (1995). *Content analysis of verbal behavior: New findings and clinical applications.* Lawrence Erlbaum Associates, Inc.

Gottschalk, L. A., & Gleser, G. C. (1979). *The measurement of psychological states through the content analysis of verbal behavior*. University of California Press.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greenberg, L. S. (1986). Change process research. *Journal of Consulting and Clinical Psychology*, *54*(1), 4. https://doi.org/10.1037/0022-006X.54.1.4

Greenberg, L. S. (2007). A guide to conducting a task analysis of psychotherapeutic change. *Psychotherapy Research*, *17*(1), 15–30. https://doi.org/10.1080/10503300600720390

Grencavage, L. M., & Norcross, J. C. (1990). Where Are the Commonalities Among the Therapeutic Common Factors? *Professional Psychology: Research and Practice*, *21*(5),

372–378. https://doi.org/10.1037/0735-7028.21.5.372

Griffiths, M. (2005). A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance Use*, *10*(4), 191–197. https://doi.org/10.1080/14659890500114359

Haight, B. K., & Webster, J. D. (1995). *The art and science of reminiscing: Theory, research, methods, and applications*. Taylor & Francis. https://doi.org/10.4324/9780203782347

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Halfon, S., Aydın Oktay, E., & Salah, A. A. (2016). Assessing affective dimensions of play in psychodynamic child psychotherapy via text analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-46843-3_2

Halfon, S., Fişek, G., & Çavdar, A. (2017). An empirical study of verb use as indicator of emotional access in therapeutic discourse. *Psychoanalytic Psychology*, *34*(1), 35–49. https://doi.org/10.1037/pap0000081

Hammer, J. H., Parent, M., & Spiker, D. (2018). Global status report on alcohol and health 2018. (V. Poznyak & D. Rekve, Eds.). *Global Status Report on Alcohol*, *65*, 1–450. https://doi.org/10.1037/coud0000248

Harpham, T., & Molyneux, C. (2001). Urban health in developing countries: A review. *Progress in Development Studies*, *1*(2), 113–137. https://doi.org/10.1177/146499340100100202

Hastie, T., Tibshirani, R., & Friedman, J. H. (2016). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). Springer.

He, Q. (2013). *Text mining and IRT for psychiatric and psychological assessment* (Doctoral dissertation). University of Twente. University of Twente [Host]. https://doi.org/10.3990/1.9789036500562

He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*(3), 441–447. https://doi.org/10.1016/j.psychres.2012.01.032

He, Q., Veldkamp, B. P., Glas, C. A. W., & de Vries, T. (2017). Automated Assessment of Patients' Self-Narratives for Posttraumatic Stress Disorder Screening Using Natural Language Processing and Text Mining. *Assessment*, *24*(2), 157–172. https://doi.org/10.1177/1073191115602551

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65. https://doi.org/10.1126/science.1200970

Hill, C. E., & Corbett, M. M. (1993). A perspective on the history of process and outcome research in counseling psychology. *Journal of Counseling Psychology*, *40*(1), 3. https://doi.org/10.1037/0022-0167.40.1.3

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), *Bergin and garfield's handbook of psychotherapy and behavior change* (pp. 87–135). John Wiley & Sons.

Himmel, W., Reincke, U., & Michelmann, H. W. (2009). Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. *Journal of Medical Internet Research*, *11*(3), e25. https://doi.org/10.2196/jmir.1123

Hirschberg, J., & Manning, C. D. (2015). Review: Advances in natural language processing. *Science*, *349*(6245), 261–266.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, *43*(3), 524–527. https://doi.org/10.1016/j.jrp.2009.01.006

Ho Yu, C., Jannasch-Pennell, A., DiGangi, S., Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, *16*(3), 730–744.

Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., & Szolovits, P. (2017). Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations. *IEEE Journal of Biomedical and Health Informatics*, *21*(5), 1449–1459. https://doi.org/10.1109/JBHI.2016.2601123

Horvath, A. O., & Luborsky, L. (1993). The role of the therapeutic alliance in psychotherapy. *Journal of Consulting and Clinical Psychology*, *61*(4), 561.

Howes, C., Purver, M., & McCabe, R. (2014). Linguistic Indicators of Severity and Progress in Online Text-based Therapy for Depression. *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 7–16.

Howes, C., Purver, M., McCabe, R., Healey, P. G. T., & Lavelle, M. (2012). Helping the medicine go down: Repair and adherence in patient-clinician dialogues. *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, 155.

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*(2), 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hoyle, R. H., Georgesen, J. C., & Webster, J. M. (2001). Analyzing data from individuals in groups: The past, the present, and the future. *Group Dynamics: Theory, Research, and Practice*, *5*(1), 41. https://doi.org/10.1037/1089-2699.5.1.41

Huang, F. L. (2018). Multilevel Modeling Myths. *School Psychology Quarterly*, *33*(3), 492–499. https://doi.org/10.1037/spq0000272

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up

the evaluation of patient-provider interactions. *Psychotherapy*, *51*(1), 10.1037/a0036841.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Ji, S., Yu, C. P., Fung, S.-f., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity, special issue: Social Big Data: Mining, Applications, and Beyond*. https://doi.org/10.1155/2018/6157249

Johnson, M. (2009). How the Statistical Revolution Changes (Computational) Linguistics. *Proceedings of the {EACL} 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* 3–12. http://www.aclweb.org/anthology/W09-0103

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of twitter data in the aftermath of violence on college campuses. *Psychological Methods*, *21*(4), 526–541. https://doi.org/10.1037/met0000099

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jurafsky, D., & Martin, J. H. (2014). Some Brief History. *Speech and language processing* (2nd ed., pp. 9–14). Pearson Education.

Jurafsky, D., & Martin, J. H. (2017). *Speech and Language Processing: An introduction to natural language processing* (3rd ed.). Pearson Prentice Hall. https://doi.org/10.1162/089120100750105975

Kamps, J., Monz, C., de Rijke, M., & Sigurbjörnsson, B. (2004). Language-dependent and language-independent approaches to cross-lingual text retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3237*, 152–165. https://doi.org/10.1007/978-3-540-30222-3_14

Karpenko, O., & Kostyuk, G. (2018). Community-based mental health services in Russia: past, present, and future. https://doi.org/10.1016/S2215-0366(18)30263-3

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., Berger, T., Botella, C., Breton, J. M., Carlbring, P., Christensen, H., de Graaf, E., Griffiths, K., Donker, T., Farrer, L., Huibers, M. J., Lenndin, J., Mackinnon, A.,

Meyer, B., … Cuijpers, P. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: An 'individual patient data' meta-analysis. *Psychological Medicine*, *45*(13), 2717–2726. https://doi.org/10.1017/S0033291715000665

Kazdin, A. E., & Nock, M. K. (2003). Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. *Journal of Child Psychology and Psychiatry*, *44*(8), 1116–1129. https://doi.org/10.1111/1469-7610.00195

Kelders, S. M., Kok, R. N., Ossebaard, H. C., & van Gemert-Pijnen, J. E. (2012). Persuasive system design does matter: A systematic review of adherence to web-based interventions. *Journal of Medical Internet Research*, *14*(6), e152. https://doi.org/10.2196/jmir.2104

Kenny, D. A., & Hoyt, W. T. (2009). Multiple levels of analysis in psychotherapy research. *Psychotherapy Research*, *19*(4-5), 462–468. https://doi.org/10.1080/10503300902806681

Kenny, D. A., & Judd, C. M. (1986). Consequences of Violating the Independence Assumption in Analysis of Variance. *Psychological Bulletin*, *99*(3), 422–431. https://doi.org/10.1037/0033-2909.99.3.422

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 233–265). McGraw-Hill.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*(1), 126–137. https://doi.org/10.1037/0022-3514.83.1.126

Kent, D. M., & Hayward, R. A. (2007). Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *Journal of the American Medical Association*, *298*(10), 1209–1212. https://doi.org/10.1001/jama.298.10.1209

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining Insights From Social Media Language: Methodologies and Challenges. *Psychological Methods*, *21*(4), 507–525. https://doi.org/10.1037/met0000091

Khazaie, H., Rezaie, L., Shahdipour, N., & Weaver, P. (2016). Exploration of the reasons for dropping out of psychotherapy: A qualitative study. *Evaluation and Program Planning*, *56*, 23–30. https://doi.org/10.1016/j.evalprogplan.2016.03.002

Kish, L. (1965). *Survey sampling*. John Wiley; Sons, Inc.

Klotzke, K., & Fox, J.-P. (2019a). Bayesian Covariance Structure Modelling of Responses and Process Data. *Frontiers in Psychology*, *10*, 1675. https://doi.org/10.3389/fpsyg.2019.01675

Klotzke, K., & Fox, J.-P. (2019b). Modeling Dependence Structures for Response Times in a Bayesian Framework. *Psychometrika*, *84*(3), 649–672. https://doi.org/10.1007/s11336-019-09671-8

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*(3), 281–299. https://doi.org/10.1037/a0020137

Knobloch-Fedders, L. M., Elkin, I., & Kiesler, D. J. (2015). Looking back, looking forward: A historical reflection on psychotherapy process research. *Psychotherapy Research*, *25*(4), 383–395. https://doi.org/10.1080/10503307.2014.906764

Korbmacher, J. M. (2014). Interviewer effects on respondents' willingness to provide blood samples in SHARE. *Working Paper Series 20-2014*, 1–13. https://doi.org/10.6103/SHARE.w4.111

Korte, J. (2012). *The stories we live by : the adaptive role of reminiscence in later life* (Doctoral dissertation). University of Twente. Enschede, The Netherlands. https://doi.org/10.3990/1.9789036534574

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. https://doi.org/10.1037/met0000105

Kothalkar, P., Rudolph, J., Dollaghan, C., McGlothlin, J., Campbell, T., & Hansen, J. H. (2018). Fusing Text-dependent Word-level i-Vector Models to Screen 'at Risk' Child Speech. *Interspeech*, 1681–1685. https://doi.org/10.21437/Interspeech.2018-1465

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Krannitz, M. A., Grandey, A. A., Liu, S., & Almeida, D. A. (2015). Workplace surface acting and marital partner discontent: Anxiety and exhaustion spillover mechanisms. *Journal of Occupational Health Psychology*, *20*(3), 314. https://doi.org/10.1037/a0038763

Krishnamurthy, P., Khare, A., Klenck, S. C., & Norton, P. J. (2015). Survival modeling of discontinuation from psychotherapy: A consumer decision-making perspective. *Journal of Clinical Psychology*, *71*(3), 199–207. https://doi.org/10.1002/jclp.22122

Krstić, M. (2019). Reading between the lines Detecting emotion-abstract language use in the web-based treatment "Look at your drinking". *Unpublished masterthesis: University of Twente*, 112.

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling* (2nd ed.). Springer New York LLC.

Kutner, M. H., & Brogan, D. R. (1982). Comparative Analyses of Pretest-Posttest Research Designs. *The American Statistician*, *34*(4), 229–232. https://doi.org/10.2307/2684066

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). {lmerTest}

Package: Tests in Linear Mixed Effects Models. https://doi.org/10.18637/jss.v082.i13

Kypri, K., Langley, J. D., Saunders, J. B., & Cashell-Smith, M. L. (2007). Assessment may conceal therapeutic benefit: Findings from a randomized controlled trial for hazardous drinking. *Addiction*, *102*(1), 62–70. https://doi.org/10.1111/j.1360-0443.2006.01632.x

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In M. Hersen, W. Sledge, A. M. Gross, J. Kay, B. Rounsaville, & W. W. Tryon (Eds.), *Encyclopedia of psychotherapy* (4th ed., pp. 143–189). John Wiley & Sons.

Lamers, S. M. A., Bohlmeijer, E. T., Korte, J., & Westerhof, G. J. (2015). The efficacy of life-review as online-guided self-help for adults: A randomized trial. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *70*(1), 24–34. https://doi.org/10.1093/geronb/gbu030

Langfred, C. W. (2007). The Downside of Self-Management: A Longitudinal Study of the Effects of Conflict on Trust, Autonomy, and Task Interdependence in Self-Managing Teams. *Academy of Management Journal*, *50*(4), 885–900. https://doi.org/10.5465/amj.2007.26279196

Larsen, R. J., & Marx, M. L. (2012). *An Introduction to Mathematical Statistics and Its Applications* (5th ed.). Pearson Education.

Lasseter, J. (1995). Toy Story.

Lau, P. L., Jaladin, R. A. M., Abdullah, H. S., Li, P., Lau, Jaladin, R. A. M., & Abdullah, H. S. (2013). Understanding the Two Sides of Online Counseling and their Ethical and Legal Ramifications. *Procedia - Social and Behavioral Sciences*, *103*(1), 1243–1251. https://doi.org/10.1016/J.SBSPRO.2013.10.453

Lepper, G., & Mergenthaler, E. (2005). Psychotherapy Research Exploring group process. *Psychotherapy Research*, *15*(4), 433–444. https://doi.org/10.1080/10503300500091587

Lepper, G., & Mergenthaler, E. (2007). Therapeutic collaboration: How does it work? *Psychotherapy Research*, *17*(5), 576–587. https://doi.org/10.1080/10503300601140002

Libbertz-Mohr, L. B. (2020). Can text mining enrich demographic data on dropout cases in an online alcohol intervention? *Unpublished masterthesis: University of Twente*.

Liberman, M. (2002). Emotional Prosody Speech and Transcripts.

Liehr, P., Marcus, M. T., Carroll, D., Granmayeh, L. K., Cron, S. G., & Pennebaker, J. W. (2010). Linguistic analysis to assess the effect of a mindfulness intervention on self-change for adults in substance use recovery. *Substance Abuse*, *31*(2), 79–85. https://doi.org/10.1080/08897071003641271

Lin, T. Y. (1983). Mental health in the third world. *Journal of Nervous and Mental Disease*, *171*(2), 71–78. https://doi.org/10.1097/00005053-198302000-00002

Lincoln, T. M., Rief, W., Westermann, S., Ziegler, M., Kesting, M. L., Heibach, E., & Mehl, S. (2014).

Who stays, who benefits? Predicting dropout and change in cognitive behaviour therapy for psychosis. *Psychiatry Research*, *216*(2), 198–205. https://doi.org/10.1016/j.psychres.2014.02.012

Linke, S., Murray, E., Butler, C., & Wallace, P. (2007). Internet-based interactive health intervention for the promotion of sensible drinking: Patterns of use and potential impact on members of the general public. *Journal of Medical Internet Research*, *9*(2), e10. https://doi.org/10.2196/jmir.9.2.e10

Lioma, C., & Keith van Rijsbergen, C. J. (2008). Part of speech n-grams and Information Retrieval. *Revue Francaise de Linguistique Appliquee*, *13*(1), 9–22. https://doi.org/10.3917/rfla.131.0009

Lo Verde, R., Sarracino, D., & Vigorelli, M. (2012). Therapeutic cycles and referential activity in the analysis of the therapeutic process. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *15*(1), 22–31. https://doi.org/10.4081/ripppo.2012.99

Loeys, T., & Molenberghs, G. (2013). Modeling actor and partner effects in dyadic data when outcomes are categorical. *Psychological Methods*, *18*(2), 220–236. https://doi.org/10.1037/a0030640

Longford, N. T. (1995). Random Coefficient Models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 519–570). Plenum Press. https://doi.org/10.1007/978-1-4899-1292-3_10

Lord, S. P., Sheng, E., Imel, Z. E., Baer, J., & Atkins, D. C. (2015). More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behavior Therapy*, *46*(3), 296–303. https://doi.org/10.1016/j.beth.2014.11.002

Losada, D. E., & Parapar, J. (2017). Psychological Features for Automatic Text Summarization. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, *25*(2), 129–149. https://doi.org/10.1142/S0218488517400153

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Maity, T. K., & Pal, A. K. (2013). Subject specific treatment to neural networks for repeated measures analysis. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*, 60–65.

Malandrakis, N., Potamianos, A., Iosif, E., & Narayanan, S. (2013). Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(11), 2379–2392. https://doi.org/10.1109/TASL.2013.2277931

Mani, I. (2012). Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, *5*(3), 1–142. https://doi.org/10.2200/S00459ED1V01Y201212HLT018

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing* (1st ed.). The MIT Press.

Marcus, D. K., Kashy, D. A., & Baldwin, S. A. (2009). Studying Psychotherapy Using the One-With-Many Design: The Therapeutic Alliance as an Exemplar. *Journal of Counseling Psychology*, *56*(4), 537–548. https://doi.org/10.1037/a0017291

Mariani, R., Maskit, B., Bucci, W., & De Coro, A. (2013). Linguistic measures of the referential process in psychodynamic treatment: The English and Italian versions. *Psychotherapy Research*, *23*(4), 430–447. https://doi.org/10.1080/10503307.2013.794399

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. https://doi.org/10.1016/J.CSI.2012.09.004

Martinez, A. R. (2010). Natural Language Processing. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 352–357. https://doi.org/10.1002/wics.76

Martinez, A. R., & Martinez, W. (2015). At the interface of computational linguistics and statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(4), 258–274. https://doi.org/10.1002/wics.1353

Maskit, B., Bucci, W., & Murphy, S. (2015). A Computer Program for Tracking the Evolution of a Psychotherapy Treatment. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, (1), 134–145. https://doi.org/10.3115/v1/W15-1216

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McCarthy, K. L., Caputi, P., & Grenyer, B. F. (2017). Significant change events in psychodynamic psychotherapy: Is cognition or emotion more important? *Psychology and Psychotherapy: Theory, Research and Practice*, *90*(3), 377–388. https://doi.org/10.1111/papt.12116

McCarthy, K. L., Mergenthaler, E., & Grenyer, B. F. (2014). Early in-session cognitive-emotional problem-solving predicts 12-month outcomes in depression with personality disorder. *Psychotherapy Research*, *24*(1), 103–115. https://doi.org/10.1080/10503307.2013.826834

McCarthy, K. L., Mergenthaler, E., Schneider, S., & Grenyer, B. F. (2011). Psychodynamic change in psychotherapy: Cycles of patient-therapist linguistic interactions and interventions. *Psychotherapy Research*, *21*(6), 722–731. https://doi.org/10.1080/10503307.2011.615070

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models Generalized, Linear, and Mixed Models* (2nd ed.). John Wiley & Sons, Ltd. https://doi.org/10.1198/tech.2003.s13

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0471721182

McManus, S., Bebbington, P., Jenkins, R., & Brugha, T. (2016). *Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. NHS Digital.

McMurran, M. (1994). *The Psychology Of Addiction* (1st ed.). Routledge. https://doi.org/10.4324/9780203401477

McNaughton, R., & Yamada, H. (1960). Regular Expressions and State Graphs for Automata. *IRE Transactions on Electronic Computers*, *EC-9*(1), 39–47. https://doi.org/10.1109/TEC.1960.5221603

Meier, P. S., Warde, A., & Holmes, J. (2018). All drinking is not equal: How a social practice theory lens could enhance public health research on alcohol and other health behaviours. *Addiction*, *113*(2), 206–213. https://doi.org/10.1111/add.13895

Melville, K. M., Casey, L. M., & Kavanagh, D. J. (2010). Dropout from internet-based treatment for psychological disorders. *British Journal of Clinical Psychology*, *49*(4), 455–471. https://doi.org/10.1348/014466509X472138

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*, *64*(6), 1306–1315. https://doi.org/10.1037/0022-006x.64.6.1306

Mergenthaler, E., & Kachele, H. (1996). Applying multiple computerized text-analytic measures to single psychotherapy cases. *Journal of Psychotherapy Practice and Research*, *5*(4), 307–317. https://doi.org/10.1155/2015/479615

Messer, S. B., & Wampold, B. E. (2002). Let's face facts: Common factors are more potent than specific therapy ingredients. *Clinical Psychology: Science and Practice*, *9*(1), 21–25. https://doi.org/10.1093/clipsy/9.1.21

Miller, W. R., & Mount, K. A. (2001). A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? *Behavioural and Cognitive Psychotherapy*, *29*(4), 457–471. https://doi.org/10.1017/S1352465801004064

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2008). *Manual for the Motivational Interviewing Skill Code (MISC)* (tech. rep.). https://doi.org/10.1017/CBO9781107415324.004

Miller, W. R., & Rollnick, S. (2012). *Motivational Interviewing, Helping People Change* (3rd ed.). Guilford Press.

Miller, W. R., & Rose, G. S. (2010). Toward a theory of motivational interviewing. *The American psychologist*, *64*(6), 527–37. https://doi.org/10.1037/a0016830

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. https://doi.org/10.1016/C2010-0-66188-8

Molenberghs, G., & Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, *61*(1), 22–27. https://doi.org/10.1198/016214505000000024

Molenberghs, G., & Verbeke, G. (2011). A note on a hierarchical interpretation for negative variance components. *Statistical Modelling*, *11*(5), 389–408. https://doi.org/10.1177/1471082X1001100501

Moritz, S., Schröder, J., Meyer, B., & Hauschildt, M. (2013). The more it is needed, the less it is wanted: attitudes toward face-to-face intervention among depressed patients undergoing online treatment. *Depression and Anxiety*, *30*(2), 157–167. https://doi.org/10.1002/da.21988

Mörtl, K., & Gelo, O. C. G. (2015). Qualitative methods in psychotherapy process research. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research* (pp. 381–428). Springer. https://doi.org/Gelo10.1007/978-3-7091-1382-0_20

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, *28*(1), 19–26. https://doi.org/10.1016/j.jsat.2004.11.001

Mulder, J., & Fox, J.-P. (2019). Bayes Factor Testing of Multiple Intraclass Correlations. *Bayesian Analysis*, *14*(2), 521–552. https://doi.org/10.1214/18-ba1115

Muñoz, R., & Montoyo, A. (2007). Advances on natural language processing. *Data & Knowledge Engineering*, *61*(3), 403–405. https://doi.org/10.1016/j.datak.2006.06.008

Muntigl, P., & Horvath, A. (2005). Language, Psychotherapy and Client Change: An Interdisciplinary Perspective. In R. Wodak & P. A. Chilton (Eds.), *A new agenda in (critical) discourse analysis* (pp. 213–240). John Benjamins.

Murphy, S. M., Maskit, B., & Bucci, W. (2015). Putting Feelings into Words: Cross-Linguistic Markers of the Referential Process. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal Putting Feelings into Words: Cross-Linguistic Markers of the Referential Process*, 80–88.

Muthén, B. (1997). Latent Variable Growth Modeling with Multilevel Data. *Sociological Methodology*, *27*(1), 453–480. https://doi.org/10.1007/978-1-4612-1842-5_7

Mykowiecka, A., Marciniak, M., & Kupść, A. (2009). Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, *42*(5), 923–936. https://doi.org/10.1016/J.JBI.2009.07.007

Nelder, J. A. (1954). The Interpretation of Negative Components of Variance. *Biometrika*, *41*(3), 544–548.

Nielsen, N. M., Smink, W. A. C., & Fox, J.-P. (2020). Small and Negative Correlations Among Clustered Observations: Limitations of the Linear Mixed Effects Model. *Manuscript accepted by Behaviormetrika*.

Nissen-Lie, H. A., Monsen, J. T., Rønnestad, M. H., Trygve Monsen, J., & Rønnestad, M. H. (2010). Therapist predictors of early patient-rated working alliance: A multilevel approach. *Psychotherapy Research*, *20*(6), 627–646. https://doi.org/10.1080/10503307.2010.497633

Nitti, M., Ciavolino, E., Salvatore, S., & Gennaro, A. (2010). Analyzing psychotherapy process as intersubjective sensemaking: An approach based on discourse analysis and neural networks. *Psychotherapy Research*, *20*(5), 546–563. https://doi.org/10.1080/10503301003641886

Nock, M. K. (2003). Progress review of the psychosocial treatment of child conduct problems. *Clinical Psychology: Science and Practice*, *10*(1), 1–28. https://doi.org/10.1093/clipsy.10.1.1

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., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of Clinical Psychology*, *67*(2), 127–132. https://doi.org/10.1002/jclp.20764

Norton, E. C., Bieler, G. S., Ennett, S. T., & Zarkin, G. A. (1996). Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *Journal of Consulting and Clinical Psychology*, *64*(5), 919. https://doi.org/10.1037/0022-006x.64.5.919

Norvig, P. (2017). On Chomsky and the two cultures of statistical learning. In W. Pietsch, J. Wernecke, & M. Ott (Eds.), *Berechenbarkeit der welt? philosophie und wissenschaft im zeitalter von big data* (pp. 61–83). Springer Fachmedien. https://doi.org/10.1007/978-3-658-12153-2_3

Novak, M., & Pahor, M. (2017). Using a multilevel modelling approach to explain the influence of economic development on the subjective well-being of individuals. *Economic Research-Ekonomska Istrazivanja*, *30*(1), 705–720. https://doi.org/10.1080/1331677X.2017.1311229

Oh, H., Rizo, C., Enkin, M., Jadad, A., Powell, J., & Pagliari, C. (2005). What is eHealth (3): a systematic review of published definitions. *Journal of Medical Internet Research*, *7*(1), e1. https://doi.org/10.2196/jmir.7.1.e1

Oliveira, I. R. C., Demétrio, C. G. B., Dias, C. T. S., Molenberghs, G., & Verbeke, G. (2017). Negative variance components for non-negative hierarchical data with correlation, over-, and/or underdispersion. *Journal of Applied Statistics*, *44*(6), 1047–1063. https://doi.org/10.1080/02664763.2016.1191624

Orlinsky, D. E., Ronnestad, M. H., & Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M. J. Lambert (Ed.), *Handbook of psychotherapy and behavior change*

(5th ed., pp. 307–389). John Wiley & Sons, Ltd.

Owen, J. (2013). Early career perspectives on psychotherapy research and practice: Psychotherapist effects, multicultural orientation, and couple interventions. *Psychotherapy*, *50*(4), 496–502. https://doi.org/10.1037/a0034617

Pasupathi, M., & Carstensen, L. L. (2003). Age and emotional experience during mutual reminiscing. *Psychology and Aging*, *42*(5), 798–808. https://doi.org/10.1037/0882-7974.18.3.430

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118. https://doi.org/10.1037/h0024436

Pedersen, E. R., Marshall, G. N., & Schell, T. L. (2016). Study protocol for a web-based personalized normative feedback alcohol intervention for young adult veterans. *Addiction science & clinical practice*, *11*(1), 6.

Peng, D., Wang, Z., & Xu, Y. (2020). Challenges and opportunities in mental health services during the COVID-19 pandemic. *General Psychiatry*, *33*(5).

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*, 162–166. https://doi.org/10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*. https://doi.org/10.15781/T29G6Z

Pennebaker, J. W., & Chung, C. (2011). Expressive writing and its links to mental and physical health. In H. S. Friedman (Ed.), *The oxford handbook of health psychology* (pp. 417–437). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195342819.001.0001

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2015). Linguistic inquiry and word count: LIWC 2015. *Mahway: Lawrence Erlbaum Associates*, *71*.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W., & Seagal, J. D. (1999). Forming a story: The health benefits of narrative. *Journal of clinical psychology*, *55*(10), 1243–1254. https://doi.org/10.1002/(SICI)1097-4679(199910)55:10<1243::AID-JCLP6>3.0.CO;2-N

Peräkylä, A. (2012). Conversation analysis in psychotherapy. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 551–574). Wiley-Blackwell. https://doi.org/10.1002/9781118325001.ch27

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., & An, L. (2017). Understanding and Predicting Empathic Behavior in Counseling Therapy. *Proceedings of the 55th Annual Meeting of the Association for Computa-*

tional Linguistics, 1426–1435. https://doi.org/10.18653/v1/P17-1131

Pfäfflin, F., Böhmer, M., Cornehl, S., & Mergenthaler, E. (2005). What happens in therapy with sexual offenders? A model of process research. *Sexual Abuse: Journal of Research and Treatment*, *17*(2), 141–151. https://doi.org/10.1007/s11194-005-4601-2

Podsiadlowski, A., Gröschke, D., Kogler, M., Springer, C., & van der Zee, K. (2013). Managing a culturally diverse workforce: Diversity perspectives in organizations. *International Journal of Intercultural Relations*, *37*(1), 159–175. https://doi.org/10.1016/j.ijintrel.2012.09.001

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Postel, M. G. (2011). *Well Connected: Web-based treatment for problem drinkers* (Doctoral dissertation). Radboud Universiteit. Nijmegen. https://repository.ubn.ru.nl/handle/2066/91233

Postel, M. G., de Haan, H. A., & de Jong, C. A. J. (2008). E-therapy for mental health problems: a systematic review. *Telemedicine and e-Health*, *14*(7), 707–714. https://doi.org/10.1089/tmj.2007.0111

Postel, M. G., de Haan, H. A., & de Jong, C. A. J. (2010). Evaluation of an e-therapy program for problem drinkers: a pilot study. *Substance Use & Misuse*, *45*(12), 2059–2075. https://doi.org/10.3109/10826084.2010.481701

Postel, M. G., de Haan, H. A., ter Huurne, E. D., Becker, E. S., & de Jong, C. A. J. (2010). Effectiveness of a web-based intervention for problem drinkers and reasons for dropout: randomized controlled trial. *Journal of Medical Internet research*, *12*(4), e68. https://doi.org/10.2196/jmir.1642

Postel, M. G., de Haan, H. A., ter Huurne, E. D., van der Palen, J., Becker, E. S., & de Jong, C. A. (2011). Attrition in web-based treatment for problem drinkers. *Journal of medical Internet research*, *13*(4), e117. https://doi.org/10.2196/jmir.1811

Postel, M. G., de Jong, C. A. J., & de Haan, H. A. (2005). Does e-therapy for problem drinking reach hidden populations? *American Journal of Psychiatry*, *162*(12), 2393.

Pryseley, A., Tchonlafi, C., Verbeke, G., & Molenberghs, G. (2011). Estimating negative variance components from Gaussian and non-Gaussian data: A mixed models approach. *Computational Statistics and Data Analysis*, *55*(2), 1071–1085. https://doi.org/10.1016/j.csda.2010.09.002

Python Software Foundation. (2020). Python Language Reference. http://www.python.org/

Que, J., Lu, L., & Shi, L. (2019). Development and challenges of mental health in China. https://doi.org/10.1136/gpsych-2019-100053

R Core Team. (2020). R: A Language and Environment for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Rao, C. R. (1973). *Linear Statistical Inference and its Applications: Second Editon* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316436

Raudenbush, S. W. (2001). Comparing Personal Trajectories and Drawing Causal Inferences from Longitudinal Data. *Annual Review of Psychology*, *52*(1), 501–525. https://doi.org/10.1146/annurev.psych.52.1.501

Raudenbush, S. W., & Bryk, A. S. (2002a). Applications in the Study of Individual Change. *Hierarchical linear models: Applications and data analysis methods* (2nd ed., pp. 160–204). SAGE Publications Ltd.

Raudenbush, S. W., & Bryk, A. S. (2002b). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Sage Publications.

Rehm, J., Shield, K. D., Gmel, G., Rehm, M. X., & Frick, U. (2013). Modeling the impact of alcohol dependence on mortality burden and the effect of available treatment interventions in the European Union. *European Neuropsychopharmacology*, *23*(2), 89–97. https://doi.org/10.1016/j.euroneuro.2012.08.001

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research [R package version 1.8.12]. https://doi.org/lity-project.org/r/psychhttp://personality-project.org/r/psych-m

Rochlen, A. B., Zack, J. S., & Speyer, C. (2004). Online Therapy: Review of Relevant Definitions, Debates, and Current Empirical Support. *Journal of Clinical Psychology*, *60*(3), 269–283. https://doi.org/10.1002/jclp.10263

Rogers, C. R. (1961). *On becoming a person: A therapist's view of psychotherapy*. Constable.

Rogers, M. A. A., Lemmen, K., Kramer, R., Mann, J., & Chopra, V. (2017). Internet-delivered health interventions that work: systematic review of meta-analyses and evaluation of website availability. *Journal of Medical Internet Research*, *19*(3), e90. https://doi.org/10.2196/jmir.7111

Rooke, S., Thorsteinsson, E., Karpin, A., Copeland, J., & Allsop, D. (2010). Computer-delivered interventions for alcohol and tobacco use: a meta-analysis. *Addiction*, *105*(8), 1381–1390. https://doi.org/10.1111/j.1360-0443.2010.02975.x

Rosner, B., & Grove, D. (1999). Use of the Mann–Whitney U-test for clustered data. *Statistics in medicine*, *18*(11), 1387–1400. https://doi.org/10.1002/(SICI)1097-0258(19990615)18:11<1387::AID-SIM126>3.0.CO;2-V

Salton, G. (1971). *The SMART retrieval system-experiments in automatic document processing*. Prentice Hall. https://doi.org/10.1109/TPC.1972.6591971

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Salvatore, S., Gelo, O. C. G., Gennaro, A., Metrangolo, R., Terrone, G., Pace, V., Venuleo, C., Venezia, A., & Ciavolino, E. (2015). An automated method of content analysis for psychotherapy research: A further validation. *Psychotherapy Research*, *27*(1), 38–50. https://doi.org/10.1080/10503307.2015.1072282

Salvatore, S., Gennaro, A., Auletta, A. F., Tonti, M., & Nitti, M. (2012). Automated method of content analysis: A device for psychotherapy process research. *Psychotherapy Research*, *22*(3), 256–273. https://doi.org/10.1080/10503307.2011.647930

Sanyal, S. (2009). The Art, Science and History of NLP. *SSRN Electronic Journal*, 1–5. https://doi.org/10.2139/ssrn.1331771

Sassaroli, S., Brambilla, R., Cislaghi, E., Colombo, R., Centorame, F., Favaretto, E., Fiore, F., Veronese, G., & Ruggiero, G. M. (2014). Emotion-abstraction patterns and cognitive interventions in a single case of standard cognitive-behavioral therapy. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *17*(2), 65–72. https://doi.org/10.7411/RP.2014.020

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019

Schaefer, D. R., Simpkins, S. D., Vest, A. E., & Price, C. D. (2011). The Contribution of Extracurricular Activities to Adolescent Friendships: New Insights Through Social Network Analysis. *Developmental Psychology*, *47*(4), 1141–1152. https://doi.org/10.1037/a0024091

Schegloff, E. (2007). *Sequence organization in interaction: Volume 1: A primer in conversation analysis* (Vol. 1). Cambridge University Press.

Schmidt, S., Zlatkin-Troitschanskaia, O., & Fox, J.-P. (2016). Pretest-Posttest-Posttest Multilevel IRT Modeling of Competence Growth of Students in Higher Education in Germany. *Journal of Educational Measurement*, *53*(3), 332–351. https://doi.org/10.1111/jedm.12115

Schrimsher, G. W., & Filtz, K. (2011). Assessment reactivity: Can assessment of alcohol use during research be an active treatment? *Alcoholism Treatment Quarterly*, *29*(2), 108–115. https://doi.org/10.1080/07347324.2011.557983

Schroder, R., Sellman, D., Frampton, C., & Deering, D. (2009). Youth retention: Factors associated with treatment drop-out from youth alcohol and other drug treatment. *Drug and Alcohol Review*, *28*(6), 663–668. https://doi.org/10.1111/j.1465-3362.2009.00076.x

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D. J., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS One*, *8*(9), 1–16. https://doi.org/10.1371/journal.pone.0073791

Schweitzer, J., & Synowiec, C. (2012). The Economics of eHealth and mHealth. *Journal of Health Communication*, *17*(1), 73–81. https://doi.org/10.1080/10810730.2011.649158

Searle, S. R. (1971). *Linear Models*. Wiley Online Library. https://doi.org/10.1002/9781118491782

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components* (3rd ed.). John Wiley & Sons. https://doi.org/10.1375/twin.14.1.25

Serrano, J. P., Latorre, J. M., Gatz, M., & Montanes, J. (2004). Life Review Therapy Using Autobiographical Retrieval Practice for Older Adults With Depressive Symptomatology. *Psychology and Aging*, *19*(2), 272–277. https://doi.org/10.1037/0882-7974.19.2.272

Shapiro, D. A. (1995). Finding out how psychotherapies help people change. *Psychotherapy Research*, *5*(1), 1–21. https://doi.org/10503309512331331106

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Smaling, A. (2003). Inductive, analogical, and communicative generalization. *International Journal of Qualitative Methods*, *2*(1), 52–67.

Smink, W. A. C., Fox, J.-P., Tjong Kim Sang, E., Sools, A. M., Westerhof, G. J., & Veldkamp, B. P. (2019). Understanding Therapeutic Change Process Research through Multilevel Modelling and Text Mining. *Frontiers in Psychology*, *10*, 1186. https://doi.org/10.3389/fpsyg.2019.01186

Smink, W. A. C., Sools, A. M., Postel, M. G., Tjong Kim Sang, E., Elfrink, A., Libbertz-Mohr, L. B., Veldkamp, B. P., & Westerhof, G. J. (2020). Analysis of the e-mails from the Dutch Web-Based Intervention 'Alcohol de Baas': Assessment of Early Indications of Drop-Out in an Online Alcohol Abuse Intervention. *Manuscript submitted for publication*.

Smink, W. A. C., Sools, A. M., Tjong Kim Sang, E., Veldkamp, B. P., & Westerhof, G. J. (2020). The Automation and Explication of Research Methods: Understanding their Interplay through a Framework, with Therapeutic Change Process Research as an Use-Case. *Manuscript submitted for publication*.

Smink, W. A. C., Sools, A. M., van der Zwaan, J. M., Wiegersma, S., Veldkamp, B. P., & Westerhof, G. J. (2019). Towards Text-Mining Therapeutic Change: A Systematic Review of Text-Based Therapeutic Change Process Research methods. *PLoS One*, *14*(12), e0225703. https://doi.org/10.1371/journal.pone.0225703

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). SAGE Publications Ltd.

Snijders, T. A., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*(4), 471–486. https://doi.org/10.1111/j.1475-6811.1999.tb00204.x

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y.-T. (2008). Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. https://doi.org//10.3115/1613715.1613751

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Sools, A. M., Smink, W. A. C., van der Zwaan, J. M., Schuffelen, P. C. J., Tjong Kim Sang, E., de Vries, B. L., Veldkamp, B. P., & Westerhof, G. J. (2019). Text Mining Research of Psychotherapeutic Processes: A State-of-the-Art Review. *Manuscript submitted for publication*.

Stark, M. J. (1992). Dropping out of substance abuse treatment: A clinically oriented review. *Clinical Psychology Review*, *12*(1), 93–116. https://doi.org/10.1016/0272-7358(92)90092-M

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, *43*(2), 476–493. https://doi.org/10.1093/ije/dyu038

Stiles, W. B., Elliott, R., Llewelyn, S. P., Firth-Cozens, J. A., Margison, F. R., Shapiro, D. A., & Hardy, G. E. (1990). Assimilation of problematic experiences by clients in psychotherapy. *Psychotherapy*, *27*(3), 411–420. https://doi.org/10.1037/0033-3204.27.3.411

Stiles, W. B., Morrison, L. A., Harper, H., Shapiro, D. A., Haw, S. K., Firth-Cozens, J. A., Harper, H., Shapiro, D. A., & Firth-Cozens, J. A. (1991). Longitudinal study of assimilation in exploratory psychotherapy. *Psychotherapy*, *28*(2), 195–206. https://doi.org/10.1037/0033-3204.28.2.195

Stivers, T. (2015). Coding Social Interaction: A Heretical Approach in Conversation Analysis? *Research on Language and Social Interaction*, *48*(1), 1–19. https://doi.org/10.1080/08351813.2015.993837

Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient Education and Counseling*, *74*(3), 295–301. https://doi.org/10.1016/j.pec.2008.11.015

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547–559. https://doi.org/10.1037/a0028226

Tanaka, H., Sakti, S., Neubig, G., Toda, T., & Nakamura, S. (2014). Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. *Proceedings of the Workshop on Computational Linguistics and Clini-*

cal Psychology: From Linguistic Signal to Clinical Reality, 88–96.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., Smyth, P., & Srikumar, V. (2015). Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 71–79. https://doi.org/10.3115/v1/W15-1209

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, *65*(1), 43–50.

Tasca, G. A., & Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychotherapy Research*, *19*(4-5), 429–437. https://doi.org/10.1080/10503300802641444

Tasca, G. A., Sylvestre, J., Balfour, L., Chyurlia, L., Evans, J., Fortin-Langelier, B., Francis, K., Gandhi, J., Huehn, L., Hunsley, J., Joyce, A. S., Kinley, J., Koszycki, D., Leszcz, M., Lybanon-Daigle, L., Mercer, D., Ogrodniczuk, J. S., Presniak, M., Ravitz, P., … Wilson, B. (2015). What clinicians want: Findings from a psychotherapy practice research network survey. *Psychotherapy*, *52*(1), 1–11. https://doi.org/10.1037/a0038252

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thompson, W. A. (1962). The problem of of negative of variance components. *Annals of Mathematical Statistics*, *33*(3), 273–289. https://doi.org/10.1214/aoms/1177705148

Thorne, A., McLean, K. C., & Lawrence, A. M. (2004). When remembering is not enough: Reflecting on self-defining memories in late adolescence. *Journal of Personality*, *72*(3), 513–541. https://doi.org/10.1111/j.0022-3506.2004.00271.x

Tjong Kim Sang, E., de Vries, B. L., Smink, W. A. C., Veldkamp, B. P., Westerhof, G. J., & Sools, A. M. (2019). De-identification of Dutch Mental Health Data [Poster presetation].

Trani, S., Lucchese, C., Perego, R., Losada, D. E., Ceccarelli, D., & Orlando, S. (2018). SEL: A unified algorithm for salient entity linking. *Computational Intelligence*, *34*(1), 2–29. https://doi.org/10.1111/coin.12147

van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandehinste (Eds.), *Selected papers of the 17th computational linguistics in the netherlands meeting* (pp. 99–114).

van den Eijnden, R. J., Lemmens, J. S., & Valkenburg, P. M. (2016). The Social Media Disorder Scale. *Computers in Human Behavior*, *61*, 478–487. https:

//doi.org/10.1016/J.CHB.2016.03.038

van den Hazel, T. S. (2020). *Predicting Early Indicators of Dropout in Online Therapy for Problem Drinkers: Using LIWC to analyse email contact between client and counsellor* (Unpublished bachelorthesis). Unpublished bachelorthesis: University of Twente.

van der Zwaan, J. M., Leemans, I., Kuijpers, E., & Maks, I. (2015). HEEM, a complex model for mining emotions in historical text. *2015 IEEE 11th International Conference on e-Science*, 22–30.

van Voren, R. (2017). Mental health and human rights in Russia—a flawed relationship. https://doi.org/10.1016/S0140-6736(17)32402-9

van Wissen, L., & Boot, P. (2017). An Electronic Translation of the LIWC Dictionary into Dutch, 703–715.

Veldkamp, B. P. (2018). Master the data mass.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*(2), 254–262. https://doi.org/10.1111/1541-0420.00032

Verbeke, G., & Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-1-4419-0300-6

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51. https://doi.org/10.1177/0962280207081238

Vernon, M. L. (2010). A review of computer-based alcohol problem services designedfor the general public. *Journal of Substance Abuse Treatment*, *38*(3), 203–211. https://doi.org/10.1016/j.jsat.2009.11.001

Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., Charlson, F. J., Davis, A., Degenhardt, L., & Dicker, D. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *386*(9995), 743–800. https://doi.org/10.1016/S0140-6736(15)60692-4.

Voutilainen, L., Peräkylä, A., & Ruusuvuori, J. (2011). Therapeutic change in interaction: Conversation analysis of a transforming sequence. *Psychotherapy Research*, *21*(3), 348–365. https://doi.org/10.1080/10503307.2011.573509

Wagenmakers, E.-J., & Farrell, S. (2004). Systematic optimization of asphaltene molecular structure and molecular weight using the quantitative molecular representation approach. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.1021/ef300604q

Wallerstein, R. S. (2001). The generations of psychotherapy research: An overview. *Psychoanalytic Psychology*, *18*(2), 243–267. https://doi.org/10.1037/0736-9735.18.2.243

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., Bruffaerts, R., de

Girlolamo, G., de Graaf, R., Gureje, O., Maria Haro, J., Karam, E. G., Kessler, R. C., Kovess, V., Lane, M. C., Lee, S., & Elisabeth Wells, J. (2007). Worldwide Use of Mental Health Services for Anxiety, Mood, and Substance Disorders: Results from 17 Countries in the WHO World Mental Health (WMH) Surveys. *The Lancet*, *370*(9590), 841–850. https://doi.org/10.1016/S0140-6736(07)61414-7

Wasserman, L. (2010). *All of Statistics* (Vol. 100). Springer Texts in Statistics. https://doi.org/10.1007/978-0-387-21736-9

Weisner, C., & Matzger, H. (2003). Missed opportunities in addressing drinking behavior in medical and mental health services. *Alcoholism: Clinical and Experimental Research*, *27*(7), 1132–1141. https://doi.org/10.1097/01.ALC.0000075546.38349.69

Weiss, S. M., & Indurkhya, N. (1995). Rule-based Machine Learning Methods for Functional Prediction. *Journal of Artificial Intelligence Research*, *3*(1), 383–403. https://doi.org/10.1613/jair.199

Westerhof, G. J., Bohlmeijer, E. T., & McAdams, D. P. (2017). The Relation of Ego Integrity and Despair to Personality Traits and Mental Health. *The Journals of Gerontology: Series B*, *72*(3), 400–407. https://doi.org/10.1093/geronb/gbv062

Westerhof, G. J., Bohlmeijer, E. T., van Beljouw, I. M., & Pot, A. M. (2010). Improvement in personal meaning mediates the effects of a life review intervention on depressive symptoms in a randomized controlled trial. *Gerontologist*, *50*(4), 541–549. https://doi.org/10.1093/geront/gnp168

Westerhof, G. J., Bohlmeijer, E. T., & Webster, J. D. (2010). Reminiscence and mental health: a review of recent progress in theory, research and interventions. *Ageing and Society*, *30*(4), 697–721. https://doi.org/10.1017/S0144686X09990328

White, M. (2007). *Maps of narrative practice*. WW Norton & Company.

White, M., & Epston, D. (1990). *Narrative means to therapeutic ends*. WW Norton & Company.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N., Burstein, R., Murray, C. J. L., & Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, *382*(9904), 1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Williams, J. M. G., Barnhofer, T., Crane, C., Hermans, D., Raes, F., Watkins, E., & Dalgleish, T. (2007). Autobiographical Memory Specificity and Emotional Disorder. *Psychological Bulletin*, *133*(1), 122–148. https://doi.org/10.1037/0033-2909.133.1.122

Woelfle, M., Olliaro, P., & Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, *3*(10), 745–748. https://doi.org/10.1038/nchem.1149

Wynn, R., & Wynn, M. (2006). Empathy as an interactionally achieved phenomenon in psychotherapy Characteristics of some conversational resources. *Journal of Pragmatics*, *38*(1), 1385–1397. https://doi.org/10.1016/j.pragma.2005.09.008

Xiao, B., Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. (2012). Analyzing the language of therapist empathy in Motivational Interview based psychotherapy. *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*.

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. (2016). Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework. *Current Psychiatry Reports*, *18*(5), 49. https://doi.org/10.1007/s11920-016-0682-5

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine*, *22*(22), 3527–3541. https://doi.org/10.1002/sim.1572

Yeung, W. F., Chung, K. F., Ho, F. Y. Y., & Ho, L. M. (2015). Predictors of dropout from internet-based self-help cognitive behavioral therapy for insomnia. *Behaviour Research and Therapy*, *73*, 19–24. https://doi.org/10.1016/j.brat.2015.07.008

Yoon, B., Phaal, R., & Probert, D. (2008). Morphology analysis for technology roadmapping: Application of text mining. *R&D Management*, *38*(1), 51–68. https://doi.org/10.1111/j.1467-9310.2007.00493.x

Zandberg, L. J., Rosenfield, D., Alpert, E., McLean, C. P., & Foa, E. B. (2016). Predictors of dropout in concurrent treatment of posttraumatic stress disorder and alcohol dependence: Rate of improvement matters. *Behaviour Research and Therapy*, *80*, 1–9. https://doi.org/10.1016/j.brat.2016.02.005

What
Works
When
for
Whom

De thesis in zeven stellingen

In deze thesis komen zeven stellingen naar voren (zie Tabel 9.1 voor een overzicht in het Engels). Uitgangspunt is één van de centrale vragen in de psychologie: *Wat* [voor psychotherapie] *Werkt er Wanneer en voor Wie* (WWWW; *What Works When for Whom*)? Een kanttekening is dat we ons beperken tot de effecten van *psychologische* therapie (met andere woorden: '*talking therapy*'). De onderzoeksdiscipline die deze vraag probeert te beantwoorden is *Therapeutic Change Process Research* (TCPR; vrij vertaald: 'onderzoek naar therapeutische veranderprocessen').

Effectstudies, ook wel bekend als *Randomized Controlled Trails* (RCTs; gerandomiseerd onderzoek met een controlegroep), zijn in deze context veel gebruikt, maar hebben –in het licht van TCPR– ook een tekortkoming. RCTs kunnen alleen vaststellen *dat* de behandeling effect heeft gehad, maar daarmee blijft open wat nu precies het onderliggende veranderproces is geweest.

Hoe kunnen we de WWWWvraag dan wel beantwoorden? Omdat de therapeutische conversatie tussen client en counselor de basis vormt voor bijna alle vormen van psychotherapie, stelt de benadering die we hier uiteenzetten de therapeutische conversatie centraal. Omdat tekst gemakkelijker te analyseren is dan spraak, is het niet vreemd dat TCPR een lange traditie heeft in de analyse van getranscribeerde taal. Omdat er zoveel verschillende onderzoeksdisciplines zijn die met tekstdata werken (denk hierbij bijvoorbeeld aan de geesteswetenschappen, informatica en linguïstiek), zijn er ook veel verschillende methoden en benaderingen voor TCPR beschikbaar (dit is stelling 1 in Tabel 9.1). In deel I van deze thesis staan we stil bij de TCPRmethoden.

Psychotherapeutische internetinterventies worden steeds populairder (onder andere ook omdat ze beschikbaar zijn tijdens een pandemie). Er zijn interventies waarbij de client en de counselor elkaar alleen online en via de e-mail 'ontmoeten'. Omdat uit eerder onderzoek al bekend is dat dit type interventies effectief is (en e-mail daarbij het enige communicatiemiddel was), moet veel van de voor TCPR relevante informatie ook in deze e-mails zitten (stelling 2).

Onderzoekers uit verschillende disciplines gebruiken ook verschillende TCPRmodellen om met deze informatie uit e-mails aan de slag te gaan (stelling

3). In deel II van de thesis gaan we in op de verschillende modellen[1] die beschikbaar zijn voor TCPR. We maken daarbij onderscheid tussen modellen die gericht zijn op het doen van *accurate voorspellingen* en modellen die vooral gericht zijn op het creëren van *uitlegbare relaties* tussen de verschillende TCPRvariabelen die in het model zitten.

In deze thesis stellen we ons op het standpunt dat voor TCPRonderzoeksdoeleinden met name multilevelmodellen geschikt zijn. Het specifieke voordeel van multilevelmodellen is dat ze goed uitlegbaar zijn en dat TCPRinformatie uit verschillende delen van de hiërarchische structuur meegekomen kan worden in het model. Voor TCPR heeft dit meerwaarde, want data bestaat vaak uit meerdere 'lagen' aan clusters: e-mails vormen het onderste cluster, de client –die de e-mails stuurde– het tweede, en de counselor zelf is het derde cluster in een multilevelmodel (stelling 4). Het multilevelmodel is dan in staat om een duiding te geven aan iedere laag in het cluster, waardoor mogelijk wordt om –bijvoorbeeld– de effectiviteit van iedere counselor te kwantificeren.

Helaas blijken multilevelmodellen ook een tekortkoming te hebben: ze zijn niet in staat om de 'negatieve' (vrij vertaald: 'divergente') effecten die clusters kunnen hebben te modelleren (stelling 5). In deel III staan we stil bij deze divergente effecten, omdat –hoewel multilevel modellen populair zijn– dit specifieke effect nog relatief onbekend is. Multilevel modellen zijn ontworpen om de 'positieve' (vrij vertaald: 'convergente') effecten in data te modelleren, waardoor het uitgesloten is dat er nog naar de divergente effecten kan worden gekeken.

En laat nu juist deze divergente effecten relevant zijn voor TCPR: omdat therapiesucces vaak het resultaat is van een interactie tussen client en counselor, kan het zijn dat een counselor een betere klik heeft met één specifieke client. De effectiviteit van de counselor kan dus variëren over de verschillende cliënten. Het gaat hier dus inderdaad om een divergent effect (stelling 6). In de thesis brengen we naar voren dat een gedegen begrip van de heterogeniteit van de counselor is een stap in de richting van het begrijpen van de WWWWvraag.

Als multilevelmodellen niet in staat zijn divergente effecten te modelleren, hoe moet dat dan wel? We introduceren *Bayesiaanse Covariantie Structuur Matrices* (BCSM), een nieuwe benadering die wel in staat is om deze effecten –die van nature divergent zijn– onder de loep te nemen (stelling 7). We presenteren een

[1] Modelleren houdt in dat aan de hand van een aantal statistische assumpties een wiskundig model tot stand komt dat de relaties tussen verschillende variabelen tot uitdrukking brengt. De input variabele x wordt dan door het model $f(x)$ 'afgebeeld' ('mapped') op de uitkomst variabele y. De relatie $x \longrightarrow f(x) \longrightarrow y$ (zoals ook te zien is in Figuur 4.1 in hoofdstuk 4) drukt dan uit hoe van input x uitkomst y te maken is. In de thesis kijken we onder andere naar hoe verschillende tekstuele aspecten van e-mails kunnen worden gerelateerd aan de afname in depressie: zo bestuderen we de relaties tussen de variabelen in tekst (x) en de afname in depressie (y).

aantal voordelen van het BCSM en laten zien welke voordelen de benadering heeft voor TCPR door antwoord te geven op de WWWWvraag.

Conclusie

In deze thesis staan we stil bij de verschillende TCPRmethoden en -modellen die gebruikt worden om de WWWW vraag te beantwoorden. TCPR wordt vooralsnog gekenmerkt door een veelheid aan verschillende disciplines en benaderingen. Omdat we (in deel I en II) in kaart hebben gebracht welke methoden en modellen er precies gebruikt worden, zal het voor andere onderzoekers makkelijker moeten zijn om bij te dragen aan een verdere integratie van TCPR.

In deze thesis introduceren we (in deel III) ook BCSM, een statistisch model dat in staat is om de WWWW te beantwoorden. De meerwaarde van de thesis ligt dus niet alleen in het gegeven overzicht van het TCPRveld, we poneren ook een model dat een stap verder kan gaan dan al het voorgaande.

What
Works
When
for
Whom

Author's contributions

The *What Works When for Whom* project was an interdisciplinary research project that relied on expertise from psychology, computer science, and statistics. Multidisciplinary mandates collaborating, which is reflected in the author's contributions for each of the chapters.

Chapter 1. Written by Wouter Smink. Gerben Westerhof, Bernard Veldkamp, and Anneke Sools provided feedback several times.

Chapter 2. Written by Wouter Smink, with help from Anneke Sools. Gerben Westerhof assisted the literature search and review. Sytske Wiegersma, Janneke van der Zwaan, Gerben Westerhof, Anneke Sools and Bernard Veldkamp provided feedback several times throughout the process.

Chapter 3. Written by Anneke Sools, with help from Wouter Smink. Wouter Smink and Pauline Schuffelen helped with reviewing the literature. First draft was written by Janneke van der Zwaan. Erik Tjong Kim Sang, Ben de Vries, Bernard Veldkamp and Gerben Westerhof provided feedback several times throughout the process.

Chapter 4. Written by Wouter Smink. Detailed feedback was provided by Gerben Westerhof, Bernard Veldkamp, Anneke Sools and Erik Tjong Kim Sang.

Chapter 5. Written by Wouter Smink, with assistance of Auke Elfrink and Lukas Libbertz-Mohr, who were supervised as part of their bachelor- and masterthesis. Erik Tjong Kim Sang conducted the data handling in Python and pre-processed the data. Marloes Postel contributed the data and gave feedback throughout the process. Anneke Sools helped Wouter Smink with the supervision of the students and gave, together with Bernard Veldkamp and Gerben Westerhof, feedback throughout the process. The data-analyses of Auke Elfrink formed the basis of the results-section, the literature review of Lukas Libbertz-Mohr contributed greatly to the introduction and discussion.

Chapter 6. Written by Wouter Smink. Jean-Paul Fox contributed to the multilevel analyses. Erik Tjong Kim Sang conducted the data handling in Python and pre-processed the data. Gerben Westerhof contributed the data. Anneke Sools Bernard Veldkamp helped with setting up project and gave detailed feedback throughout.

Chapter 7. Based on the bachelorthesis of Natalie Nielsen, who was supervised by Wouter Smink and Jean-Paul Fox, who jointly prepared the manuscript for publication.

Chapter 8. Simulation study conducted by Wouter Smink, further data-analyses conducted by Jean-Paul Fox. Wouter Smink and Jean-Paul Fox contributed equally to the paper, but as the BCSM model was developed by Jean-Paul Fox, both authors agreed that Jean-Paul Fox could submit the paper.

Chapter 9. Written by Wouter Smink. Gerben Westerhof, Bernard Veldkamp, and Anneke Sools provided feedback several times.

List of Figures

What
Works
When
for
Whom

List of Tables

Citing? Yes please!

Smink, W. A. C. (2021). *What Works When for Whom? A methodological reflection on Therapeutic Change Process* [Doctoral dissertation, University of Twente]. https://doi.org/10.3990/1.9789036550338