THE UNIVERSITY *of* EDINBURGH

Optimising methodological practices within psychology: A parapsychology case-study

Abby Lauren Pooley

Doctor of Philosophy

The University of Edinburgh

2024

To Michael

Abstract

Parapsychology is a contentious research field within psychology. One of the most active research areas is the investigation of the psi hypothesis using the experimental paradigm known as the psi ganzfeld. The psi ganzfeld is a form of mild sensory deprivation, during which participants attempt to perceive a randomly chosen video or image target. The participant is tasked with becoming aware of a target, which may be shown in a different location with no one watching (clairvoyance design), someone watching it in a different location (telepathy design), or the target is chosen after the participant makes their decision (precognition design). During the ganzfeld, participants are asked to vocalise any impressions, sensations, and experiences, which are often audio recorded and are known as the mentation. After the session, participants rate which video or image clip most closely matches their experience. If the participant's top choice matches the target clip, the session is considered a "hit." However, due to the contentious nature of this experiment, the psi ganzfeld is an ideal case for demonstrating current methodological weaknesses and proposing recommendations that can improve research practices more broadly.

This thesis is structured around two overarching themes: the first theme deals with broader methodological issues in psychological research, using the psi ganzfeld experiment as a case study. It discusses the replication crisis in psychology, situating parapsychology within this context, and explores the methodological and statistical challenges common to both fields, such as questionable research practices and the limitations of meta-analyses. A critical examination of meta-analyses of the psi ganzfeld is provided, highlighting how different methodological decisions regarding study inclusion influence the outcomes of literature reviews. An alternative approach to conventional meta-analyses is also presented, demonstrating how to triangulate and synthesise contradictory results from multiple meta-analyses to achieve a more comprehensive understanding of research decisions and study outcomes.

The second theme of the thesis focuses on methodological improvements specific to parapsychology, which can also benefit the broader field of psychology. Experimental data from a precognition study demonstrates the flexibility in constructing statistical analyses, via specification curve analysis. Further, verbal reports from ganzfeld precognition studies are analysed using a novel quantitative language analysis, offering new insights into a previously

under-examined data source in parapsychology. By applying novel methods to secondary data, these chapters provide a more comprehensive understanding of the data. Given the controversy surrounding the psi ganzfeld task, one of the empirical chapters addresses methodological concerns by proposing an innovative approach to validating study software—an aspect not commonly addressed in either parapsychological or psychological research. The last chapter presents a meta-regression of psi telepathy studies, examining the impact of various methodological features on study outcomes, and offering fresh insights into over 35 years of research.

Overall, this thesis highlights the ongoing methodological challenges in psychology, even after the replication crisis produced numerous improvements and recommendations. By using the psi ganzfeld as a case study, it underscores the need for greater transparency regarding researchers' degrees of freedom when designing reviews, synthesising divergent meta-analyses through innovative approaches like multiverse meta-analyses and employing novel methods to gain deeper insights into existing data. The recommendations and examples drawn from parapsychology, despite its contentious nature, have broader implications for advancing methodological rigour in psychology as a whole.

Lay Summary

Parapsychology, a field that investigates phenomena beyond the current understanding of psychology, is highly debated among scientists. One of its most active areas of research is the "psi ganzfeld" experiment, which explores whether people can perceive information in ways that go beyond the known senses. In these experiments, participants are placed in a state of mild sensory deprivation and try to perceive a randomly chosen video or image. This target may be shown at a distant location without anyone watching (clairvoyance), viewed by someone elsewhere (telepathy), or selected after the participant has made their guess (precognition). During this process, participants describe their thoughts and feelings, which are recorded and later analysed. If a participant's highest rated target matches the randomly selected target, it is considered a success, or a "hit".

Due to its controversial nature, the psi ganzfeld experiment is a prime example for highlighting current weaknesses in research methods and suggesting ways to improve them. This thesis explores two main themes. The first part of the thesis uses the psi ganzfeld experiment to discuss common problems in psychological studies, like the difficulty of replicating results and inconsistent research practices. It examines how these challenges are also present in parapsychology and shows how the way researchers select studies for review can change the outcomes of those reviews. This thesis also highlights concerns that researchers have a lot of control and flexibility in their decision-making processes at every level of research.

The second part focuses on specific ways to improve research methods in parapsychology that could benefit psychology. It introduces an innovative approach to analysing participants' verbal reports during psi ganzfeld experiments, revealing patterns that were previously overlooked. It also proposes a new way to validate the software used in such studies, which is rarely discussed in either parapsychology or psychology. Finally, it presents a comprehensive review of studies on telepathy, exploring how different study designs impact results.

Overall, this thesis shows that even after the "replication crisis", which led to many improvements in psychological research, significant methodological challenges remain. By using the psi ganzfeld experiment as a case study, it emphasises the need for more transparent research practices, better ways to handle conflicting study results, and new methods for analysing existing data. It concludes that parapsychologists and psychologists need to spend

more time observing and producing better theories which will make experiments more comparable and replicable.

Acknowledgements

First, I want to express my heartfelt thanks to my many PhD friends who have supported me over the past five years: Sam, Tanvi, Espe, Emma, Joléne, and Lasse[1]. Your support, guidance, hugs, and the space to vent and ask stupid questions have been invaluable.

I am deeply grateful to the WA Skeptics for funding my PhD stipend. Without this financial support, completing this thesis would not have been possible. I also appreciate the BIAL Foundation for indirectly supporting my work through my primary supervisor, Professor Watt. This allowed me to contribute to her research project (grant number 190/18) and use its data in my thesis.

My sincere thanks go to my supervisor, Dr Aja Murray, for her support, statistical checks, and wise advice on navigating thesis projects. Her guidance has been crucial, and has taught me that, even if things go wrong, projects can still be salvaged.

I want to express my deepest gratitude to my supervisor, Professor Caroline Watt. I couldn't have asked for a more exceptional mentor and colleague. From my undergraduate dissertation project through the five years of this PhD journey, your guidance has been invaluable. This journey was far from easy, with a pandemic, the loss of a close family member, my numerous internships and jobs (sorry!), and all the other challenges life threw my way. Throughout it all, you consistently supported me and endorsed every opportunity that came my way. This thesis would not have been possible without your guidance and support. As your last student, I can only hope that this work reflects and honours your dedication and knowledge of the field, and meets the high standards you set. Now, as I move forward with a wealth of skills and experiences (and numerous strings to my bow!), all I can say is, thank you.

Finally, to Michael. I cannot define how much your support, patience and love has ensured that I completed this degree. For this, I dedicate this thesis to you.

[1] To anyone else I accidentally omitted, I sincerely apologise. However, knowing the friends I keep, I'm sure you will remind me that you were not mentioned.

I declare that this thesis is my own composition and that it has not been submitted for any other degree or professional qualification. I confirm that, except where parts of jointly authored publications have been included, the work presented is entirely my own. My contributions, as well as those of the co-authors, have been clearly indicated below and in each instance of collaborative work. I also confirm that appropriate credit has been given throughout this thesis for any references to the work of others.

Abby L. Pooley (September, 2024)

Publications and Contributions

Some of the empirical chapters of this thesis have been published, have study documents available online, or, use previously published research for novel, secondary data analysis. This section lists the relevant publication and study documentation corresponding to each chapter as well as the author contributions and my contribution to the projects.

Chapter 3: Meta-analysis of telepathy psi ganzfeld studies

This chapter has been published in the *Journal of Anomalous Experiences and Cognition*:

Pooley, A., L., Murray, A. L., & Watt, C. (2023). Understanding the Factors at Play in the Sender-Receiver Dynamic During the Telepathy Ganzfeld: A Meta Analysis. *Journal of Anomalous Experience and Cognition, 3*(1), 42–77. https://doi.org/10.31156/jaex.23878

The work presented in Chapter 3 had the following contributions:

Abby L. Pooley: Conceptualisation, literature search, data extraction, formal analysis, writing - original draft, writing - review & editing.

Caroline Watt: Conceptualisation, data extraction, writing - review & editing, supervision.

Aja L. Murray: Conceptualisation, analysis support, writing - review & editing, supervision.

Chapter 4: Methodological umbrella review of psi ganzfeld meta-analyses

Chapter 4 is a methodological umbrella review, which was originally preregistered as a specification-curve multiverse meta-analysis. However, due to misinterpretation of the method, the project was transformed into a systematic review. Thus, the specification curve multiverse meta-analysis was formally cancelled on the preregistration website (see below). The systematic review in Chapter 3 was not preregistered itself but used the 'Which' and 'How' factors outlined in the preregistration documents. The systematic review was not preregistered as data extraction had already been completed.

The preregistration is available at: https://edin.ac/3UDpfuR

The cancellation of the project is available at: https://edin.ac/3Vr7HCU

The work presented in Chapter 4 had the following contributions:

Abby L. Pooley: Conceptualisation, literature search, data extraction, data coding, formal analysis, writing - original draft, writing - reviewing and editing.

Caroline Watt: Data coding, writing - reviewing and editing, supervision.

Aja L. Murray: Conceptualisation, writing - reviewing and editing, supervision.

Chapter 5: Software Validation

Chapter 5 reports on software validation conducted on Koestler Parapsychology Unit (KPU) Study 1074. The principal investigator was my primary supervisor, Professor Caroline Watt. Professor Watt had responsibility over the study design and preregistration, recruitment and training of experimenters and coordinators; data checking and analysis. I was a co-investigator on the project, liaising with the programmers both for experiment program and software validation, I performed randomness checks, supported experimenters, assisted with data management and double-checking of data; transcription of mentation reports (Chapter 6, below) and analysis (Chapters 4,5,6). The software validation was conducted by a member of the University of Edinburgh, Dr Pawel Orzechowski, who has no affiliation with the research team. He was contracted to conduct independent software validation. The subsequent analysis and reporting of the software validation was produced by myself, independently.

The preregistration documents are available at: https://edin.ac/4eYch2R The final funding report is available at: https://edin.ac/4e8Wo8t

The software validation code is available at: https://github.com/drpawelo/study_testing_precognition

The work presented in Chapter 5 had the following contributions:

Abby L. Pooley: Conceptualisation, methodology, formal analysis, writing - original draft, writing - reviewing and editing.

Caroline Watt: Conceptualisation, methodology, writing - reviewing and editing, supervision.

Pawel Orzechowski: Software programming.

Chapter 6: Specification curve analysis of the psi ganzfeld

Chapter 6 uses data from KPU 1074, as detailed above. For this chapter, I collated the experimental data, checked the data and conducted the specification-curve analysis.

The work presented in Chapter 6 had the following contributions:

Abby L. Pooley: Conceptualisation, data curation, formal analysis, writing original draft, writing - reviewing and editing.

Caroline Watt: Writing - reviewing and editing, supervision.

Aja L. Murray: Writing - reviewing and editing, supervision.

Chapter 7: Mentation analysis

Chapter 7 uses data from KPU 1039 and KPU 1074, as detailed above. The final results from KPU 1039 were published prior to my PhD studies, however, for this chapter, I conducted secondary data analysis on the verbal reports produced by the participants, which have not been analysed previously. I transcribed the verbal reports from KPU 1039 and Wave 3 from KPU 1074: mentations from Wave 1 and 2 were transcribed by undergraduate dissertation students, under the supervision of Professor Caroline Watt.

The preregistration of KPU 1039 is available at: https://edin.ac/3WnV6QH Publication of KPU 1039 is available at: https://journals.lub.lu.se/jaex/article/view/23878

The work presented in Chapter 7 had the following contributions:

Abby L. Pooley: Conceptualisation, data extraction, data curation, formal analysis, writing - original draft, writing - reviewing and editing.

Caroline Watt: Writing - reviewing and editing, supervision.

Contents

Preface

Psi is an umbrella term used by parapsychologists which encompasses both extrasensory perception (ESP) and psychokinesis (PK). Psychokinesis involves the direct influence of the mind on a physical system, without the mediation of any known physical energy. ESP, on the other hand, refers to the acquisition of information or a response to an external event, object, or influence—whether mental or physical, and occurring in the past, present, or future—through means that do not involve any known sensory channels.

ESP is further categorised into three sub-types: telepathy, clairvoyance, and precognition. Telepathy involves acquiring information about the thoughts, feelings, or activities of another conscious being. Clairvoyance refers to obtaining hidden information about an object or contemporary physical event, with the assumption that this information comes directly from an external physical source rather than from another conscious being. Precognition involves obtaining information about a future event that cannot be deduced from normally known data in the present. At a theoretical level, it is difficult to distinguish between telepathy, clairvoyance, and precognition, as these phenomena often overlap depending on how they are tested rather than on distinct causal mechanisms. For example, a clairvoyance experiment might also be considered a telepathy experiment if the information is thought to be received from another person. Similarly, both could be regarded as forms of precognition if the information becomes known to someone in the future. More recently, the term anomalous cognition (AC) has been favoured as it does not assume paranormal origins of information transfer or perception as the mediating mechanism. Throughout this thesis, the terms extrasensory perception and anomalous cognition will be used interchangeably, without making assumptions about the mechanism of information transfer (whether paranormal or otherwise).

In ESP/AC experiments, participants are often tested in two modalities: forced-choice and free-response. In forced-choice experiments, the possible targets are finite and known to the participant in advance, such as playing cards from a standard deck. In free-response studies, however, participants attempt to identify a target that is not part of a pre-defined set but is instead selected from a much larger, often open-ended dataset—such as a collection of photographs or video clips. This format allows participants to explore their impressions and

sensations freely, as the range of potential targets is relatively unlimited. For a more nuanced discussion please refer to Cardeña (2018).

Chapter 1

Introduction

> The contentious nature of purported psi phenomena has instigated much debate and critical evaluation of the research methodologies and analyses employed. Such scrutiny has fuelled methodological refinements and developments in order to eliminate flaws that could lead to the making of a type one error...This attention to detail renders parapsychology a good area to consider potential problems in research more generally.
>
> (Holt et al., 2012, p. 94-95)

1.1 Chapter overview

Parapsychology, defined as the scientific study of purported psi phenomena (Cardeña, 2018), has a long and detailed history. This chapter briefly outlines the origins of the academic field, key terminology, and important dates, providing essential background information for the subsequent chapters. The following chapters delve into specific issues affecting parapsychological research, which also reflect broader challenges in psychology research. The main focus of this chapter is the psi ganzfeld experiment, a widely used paradigm within parapsychology and the basis for this thesis. It covers the origins of the ganzfeld design, notable debates, and key research findings. The discussion situates the ganzfeld findings within broader methodological issues in psychology, including the value and limitations of meta-analyses, questionable research practices, and concerns about methodological rigour. The chapter concludes with an overview of the empirical chapters in this thesis.

1.2 Parapsychology, psi and the ganzfeld

Since the founding of the Society for Psychical Research (SPR) in 1882, parapsychological research has endeavoured for unbiased testing of the psi hypothesis

through the scientific method. Basic scientific methodological elements such as hypothesis testing, designs created to eliminate confounding variables and, appropriate statistical analysis embody research within this field (Cardeña et al., 2015). Parapsychology has also contributed to methods and subject areas later integrated into mainstream psychology, such as the first use of randomisation, masking procedures (Hacking, 1988), comprehensive use of meta-analysis (Gupta & Agrawal, 2012) and study preregistration (Johnson, 1976; Watt, 2005). These features allowed the professional organisation of parapsychology to become an affiliate of the American Association for the Advancement of Sciences in 1969 (Cardeña et al., 2015).

The founding of organisations, such as the SPR, reflects wider interest in the paranormal, magic and superstitions, which are common beliefs held by the general population. For over a century, researchers have investigated proclaimed anomalous experiences. These include reports of apparent extrasensory perception (ESP) and psychokinesis - all commonly covered by the umbrella term, psi (Watt & Tierney, 2014). Extrasensory perception includes purported telepathy (the transfer of information between two persons, unmediated by their senses or logical inference), clairvoyance (obtaining information about a distant state of affairs) and precognition (being affected by an event taking place in the future that could not have been foreseen; Cardeña (2018).

The most recent Gallup survey conducted in the United States, with 1002 adult respondents, found 73% of Americans believe in at least one type of paranormal phenomenon, with ESP being the most popular belief at 41% (Moore, 2005). Asking about specific types of ESP, 31% of respondents believed in telepathy and 26% of respondents believed in clairvoyance. The poll showed no statistical differences across age, gender, education, race or region of the country and the overall percentage of those with at least one paranormal belief has held constant from the previous poll in 2001 (76% in 2001 vs 73% in 2005; Moore, 2005). Likewise, a UK based telephone poll conducted by Ipsos MORI of 1005 British adults, found that while 77% of the British public deny they are superstitious, half touch wood to avoid bad luck (51%). Meanwhile, three in ten of those who claim not to be superstitious cross their fingers for good luck and 41% of respondents reported they believe in telepathy (Ipsos MORI, 2007). However, the UK poll found women were more superstitious and had greater beliefs across the board than men (Ipsos MORI, 2007). Thus, beliefs about the paranormal and superstitions are common, as are reported personal experiences. Surveys of the general population have found 33-50% of individuals have

reported a telepathic experience (Targ et al., 2000) and 67% of respondents claimed to have had a psi-related experience in a large survey conducted by the University of Chicago in 1987 (Greeley, 1987). More recently, surveys have found little change in the prevalence of subjective psi experiences (Pechey & Halligan, 2011), with 48% of a British sample reporting anomalous experiences occurring "sometimes/often" (Pechey & Halligan, 2012). Therefore, beliefs and reported cases of spontaneous anomalous experiences are common within the general population, and with systematic reporting of cases since the 19th Century, it can be concluded that psi-related experiences are universal (Watt & Tierney, 2014).

1.2.1 Extrasensory Perception or Anomalous Cognition?

Psi typically falls into two major areas: 1) extrasensory perception (ESP) and 2) psychokinesis, or PK. The term ESP is an umbrella name for telepathy, clairvoyance and precognition. These are ontological distinctions that originate from the implicit assumption that anomalous cognition is quasi-perceptual process (Palmer, 2015). However, in practice it is impossible to entirely isolate these forms of ESP: in a design where information (or a target) is 'sent' by one person, one could get this information from the other's mind (telepathy), or from the information itself (clairvoyance). Or, the individual could have accessed the information as it existed at the time, or as it would exist at some future point (precognition). Telepathy designs are the most difficult to control against sensory leakage, due to necessary safe guarding of the target information and the sender (Palmer, 2015). Although telepathy is the hardest design to implement, and precognition is the simplest (as no target shielding is needed), telepathy is more plausible to participants than precognition (Palmer, 2015). Although the phrase extrasensory perception is commonly used in parapsychological research, its use has been criticised for being burdened with assumptions about the underlying mechanisms (May et al., 1995; Palmer, 1992). For example, extrasensory perception implies an underlying perceptual ability other than the usual senses - however, there is no reason to suggest that psi is not perceptual (Cardeña et al., 2015). In addition, as there is a debate about the nature of psi, it is perhaps best to use a more neutral term, such as anomalous cognition (AC; Cardeña et al., 2015). Since the mid-nineties, the phrase anomalous cognition has been increasingly used. For example, May et al. (1995) recommends the use of anomalous cognition, instead of ESP, to refer to plausible acquisition of information in ways that are currently unexplained. More specifically, explicit anomalous cognition is where individuals consciously intend to interact with other systems (e.g.,

attempting to retrieve information via telepathy), whereas implicit anomalous cognition involves no conscious awareness of the process but individuals nonetheless respond to the information, usually physiologically (Cardeña et al., 2015; May et al., 1995).

1.2.2 From case-studies, to card guessing and beyond

As stated, parapsychology is the examination of psi phenomena using scientific methods and has evolved from unverifiable first-person reports of subjective experiences, to an experimental approach (Holt et al., 2012). The early days of experimental laboratory-based research into AC used forced-choice card guessing methods, where the participant knows the target alternatives before they make a guess (Holt et al., 2012; Milton, 1997b). For example, J. B. Rhine - hailed as the popularist of experimental and statistical analysis of psi research (Mauskopf & McVaugh, 1980), and his colleague Dr Zener created the 'Zener cards'. Each pack of 25 cards are composed of five symbols (circle, square, wavy lines, star and cross) and study designs can be created to test for telepathy, clairvoyance and precognition. This allows for statistical analysis to be performed to assess if correct guessing is above, or below, chance (Holt et al., 2012). Forced-choice studies, such as card guessing, allow researchers to conduct high numbers of trials within one session, and a single study can consist of hundreds or thousands of trials (Milton, 1997b). However, they lack ecological validity and can lead to fatigue, for both the researcher and participant. Nonetheless, meta-analyses of forced-choice studies from 1935 onwards have obtained statistically significant effect sizes, indicating overall above chance scoring (Honorton & Ferrari, 1989; Storm et al., 2010a).

During the 1960s and 70s, there was a shift within psychology from behaviourism towards more humanistic paradigms. In parapsychology there was greater interest in capturing the spontaneous nature of AC within the laboratory, which was more ecologically valid and had a greater emphasis on introspection (Honorton, 1977). Free-response designs allow participants to report their stream of consciousness (mentation) with no constraints, providing richer detail about the individual's experiences. Unlike forced-choice studies, a free-response session is usually devoted to a single trial, which can last up to two hours (Milton, 1997b). Reflecting this paradigm shift, dream ESP studies were conducted in several research institutions, such as such as the Maimonides Medical Center (Ullman et al., 2002). As reported cases of AC occur more frequently in dreams, meditation, waking hallucinations

and hypnagogic states, free-response studies are used in an attempt to induce altered states of consciousness via hypnosis, sleep and sensory isolation (Holt et al., 2012).

Likewise, in the early 1970s there was growing exploration of a perceptual isolation technique, the ganzfeld. This research was driven by amounting evidence showing anomalous communications were frequently associated with internal attention states, due to reduced perceptual processing (Honorton, 1977, 1995). A common hypothesis as to why the ganzfeld may facilitate psi is that it induces a psi conducive state of consciousness, including vivid imagery and time distortions (Tart, 1978; Wackermann et al., 2008). Consistent findings from this era were obtained by Sargent and colleagues, across a number of studies with psi tasks correlating with self-reported shifts in states of consciousness (Sargent, 1980). In addition, Palmer and colleagues found a similar correlation between altered states of consciousness and performance (Palmer et al., 1979).

Research in the ganzfeld continues to the present day, with altered states of consciousness research having something of a resurgence. Parapsychologists have claimed to have found replicable evidence for anomalous cognition in the ganzfeld, especially when using selected participants (Storm et al., 2010b). However, not everyone experiences notable shifts in consciousness during the ganzfeld and individual differences in responsiveness remain unclear (Wackermann et al., 2008). Thus, exploring individual differences to ganzfeld stimulation and the potential underlying mechanisms that may explain success in psi tasks is warranted (Marcusson-Clavertz & Cardeña, 2011). Research on the ganzfeld has been meta-analysed repeatedly (Bem & Honorton, 1994; Hyman & Honorton, 1986; Milton & Wiseman, 1999; Storm et al., 2010b) and is the most consistently supportive database for psi in the past few decades (Cardeña, 2018). Nonetheless, the database is not without controversy: there is evidence to suggest that the findings are inflated due to questionable research practices (QRPs) and sub-par methodology (Bierman et al., 2016; Wagenmakers et al., 2011). Hence, some argue that "a meta-analysis is a poor substitute for one large well-conducted trial" (Green et al., 2003, p. 231).

1.2.3 A brief history of the Ganzfeld

The ganzfeld is a mild form of sensory deprivation that was originally developed by perceptual psychologists and later adopted in psi research (Bertini et al., 1964; Honorton & Harper, 1974). The term Ganzfeld translates from German as "whole field" and is commonly used in psi research (Cardeña, 2018; Holt et al., 2012). In a standard ganzfeld study, the

participant sits in a comfortable chair and listens to relaxation instructions (such as an audio recorded breathing exercise) and exposed to white or pink noise (unpatterned random frequencies, similar to radio static) via headphones, with a translucent cover over their eyes (such as an eye-mask) in front of a red light to create a homogeneous visual field (Cardeña et al., 2015; Cardeña, 2018). The participant's task is to become aware of a randomly chosen image, or video clip, which might be shown simultaneously in a different location (e.g., on a computer screen in a different room) with nobody watching it (clairvoyance design), someone watching it (telepathy design), or the clip is chosen after the participant makes their selection (precognition design; Cardeña (2018). During the ganzfeld, the participant is asked to describe any impressions or sensations they are experiencing (mentation report), which is generally audio recorded. After the experiment, the participant reviews four images, or video clips (one target plus three decoys), and ranks the similarity of their impressions with the presented targets. Thus, there is a 25% (mean chance expectation; MCE) chance that the participant will get a 'hit'.[2]

The psi ganzfeld method is based upon the noise reduction theory: according to this theory psi information is subtle and most likely to remain in the unconscious, hence reducing bodily actions and sensory information may facilitate psi (Honorton & Harper, 1974; Honorton, 1977). Thus, procedures which reduce both internal and external stimuli, such as meditation, hypnosis and the ganzfeld, should facilitate awareness of psi (Cardeña, 2018). Since the 1970s the ganzfeld has been repeated, and continues to be repeated and thus, provides a wealth of information to explore - especially via meta-analysis (Holt et al., 2012).

However, some researchers argue that there is no consensus in the scientific community regarding the evidence for psi (Cardeña et al., 2014; Milton, 1999). The analysis of other groups of studies suggests there is evidence of selective reporting (Bösch et al., 2006) and significant psi effects correlating with poor study quality (S. Schmidt et al., 2004). Nonetheless, both sides can agree that due to the inherent complexities and difficulties in drawing conclusions from spontaneous paranormal experiences, the psi question is best addressed with controlled experiments (Cardeña et al., 2014). Furthermore, the methodological quality of laboratory-based parapsychological research is high and arguably higher than other comparable areas of science, perhaps due to the controversial claims being

[2] Researchers are not restricted to a four-choice design, but it has increasingly become the norm.

tested (Cardeña et al., 2014). In the following sections, the notable chapters of the ganzfeld debate are addressed in full.

The Hyman-Honorton Debate and the Joint Communiqué

Between 1974 and 1981, 42 ganzfeld studies were conducted by 47 different investigators. Hyman (1985) and Honorton (1985) independently analysed these studies and came to different conclusions. Hyman, a more sceptical researcher, highlighted methodological weaknesses, and performed a factor analysis which revealed a significant correlation between specified flaws and the rate of success in the ganzfeld. In response, Honorton criticised Hyman's categorisations of study flaws and meta-analysed the 28 studies from the database that used direct hits as the outcome measure (Baptista et al., 2015; Honorton, 1985). Honorton reported there was evidence for psi, whereas Hyman did not. The debate resulted in the Joint Communiqué (Hyman & Honorton, 1986), with both authors agreeing that there was an overall effect in the database. However, they differed on the extent of evidence for psi. The Joint Communiqué also proposed stricter methodological guidelines, called for more experiments conducted by a wider range of experimenters following the more stringent guidelines. In particular, the paper called for stricter security precautions against sensory leakage, testing and documentation of randomisation procedures for selecting targets and target pools, statistical corrections for multiple analyses, advance specification of the experiment, checking for recording errors, full documentation of the experimental procedures and the status of statistical testing (i.e., planned or post-hoc).

Post Joint Communiqué and the Psychophysical Research Laboratories

Nearly a decade later, Bem and Honorton (1994) published a meta-analysis of 10 automated ganzfeld studies conducted at Honorton's Psychophysical Research Laboratories (PRL). The PRL studies were designed to conform with the new, more stringent methodological guidelines in the Joint Communiqué. A total of 329 sessions were conducted with an overall hit rate of 32%, a hit rate significantly different from mean chance expectation (see Table 1.1). Bem and Honorton (1994) concluded that this database showed evidence for the ganzfeld psi effect and these findings were both robust and reproducible (Baptista et al., 2015). However, Milton and Wiseman (1999) conducted a meta-analysis of 30 studies conducted post-PRL (1987 - 1997), by ten different first authors across seven laboratories, and concluded there was no evidence to support the ganzfeld psi hypothesis. The database of studies yielded a null result with an overall hit rate of 27.5% (see Table 1.1). The

authors concluded that the PRL findings had failed to be replicated and the ganzfeld paradigm did not provide evidence for psychic functioning.

Milton and Wiseman's findings triggered a debate within the parapsychology community, with many criticising their inclusion criteria and statistical analyses (Baptista & Derakhshani, 2014; Storm & Ertel, 2001). This included many questioning why a highly significant series using creative participants (Dalton, 1997) was excluded, however these findings were reported after the time frame defined by Milton and Wiseman. Nonetheless, Milton (1999) updated the database with nine more studies and confirmed there there was no support for the psi hypothesis in the psi ganzfeld database ($Z = 1.45$, $p = 0.074$) when excluding Dalton (1997). However, when this study was included in the analysis the database become significant ($Z = 2.28$, $p = 0.011$). Milton commented that the database only became significant due the Dalton study, thus there was insufficient evidence to suggest success by a range of investigators, in contrast to other findings (e.g., Bem & Honorton, 1994). Furthermore, even with the inclusion of the Dalton (1997b) data in the Milton and Wiseman (1999) database, the effect sizes reported from the PRL series (Bem & Honorton, 1994) were not replicated and were in fact significantly lower ($t(48) = 2.22$, $p = 0.016$).

Although there was a debate about Milton and Wiseman's conclusions, there was agreement about the significantly reduced overall effect size in the ganzfeld database (Baptista et al., 2015). The PRL studies found a hit rate of 32.2% (Bem & Honorton, 1994), whereas the next meta-analysis found a hit rate of 27.5% (Milton, 1999). More recently, there has been a re-analysis of the Milton and Wiseman database, suggesting that when similarity of study populations across the databases only are considered, the Milton and Wiseman database did replicate the PRL findings (Baptista & Derakhshani, 2014).

A later meta-analysis conducted by Bem et al. (2001) suggested that the effect size decrease in the replication outcome was significantly related with the degree to which the study adhered to the standard ganzfeld protocol. Bem and colleagues found ten new studies and added these to the Milton and Wiseman (1999) database. Studies that ranked above the midpoint (4.0) for standard-ness produced significant results (31.2% hit-rate, 1278 trials, 29 studies, exact binomial $p < .001$) compared to studies below the midpoint, which produced a non-significant hit-rate of 24%. This difference was significant ($U = 190.5$, $p < .05$). The 10 new ganzfeld replication studies (conducted after the Milton-Wiseman cut-off date) yielded a hit rate of 30.1% and found evidence for an overall ganzfeld effect ($Z = 2.59$, $p < .001$; Bem

et al., 2001). However, the retrospective database analysed by Bem et al. (2001) was not an update of the ganzfeld, but mainly a critique of the Milton and Wiseman database and was more comparative than comprehensive (Storm et al., 2010b). Nonetheless, Bem et al. (2001) argued that Milton and Wiseman (1999) meta-analysis included studies that did not adhere to these 'standard designs', such as those that used musical targets (Willin, 1996a, 1996b). The authors concluded studies that deviated from the original PRL protocol were less successful with an overall hit-rate of 24%, compared to an average hit-rate of 31% for studies following the protocol (Bem et al., 2001). Storm and Ertel (2001) also conducted a review in response to the Milton and Wiseman (1999) findings, however their database was criticised for including studies conducted before the Joint Communiqué recommendations (see Table 1.1). The authors argued that the combined database (11 studies pre-Joint Communiqué plus the Milton and Wiseman database) confirmed the findings by Bem and Honorton (1994) (Storm & Ertel, 2001).

Current research

In 2010, Storm et al. (2010b) conducted an updated meta-analysis of the ganzfeld. The database contained 30 studies conducted between 1997-2008, by 36 different lead investigators with a total of 1,648 trials. Their inclusion criteria was only ganzfeld studies that had more than two participants, used a random number generator (or random number table) for target selection, and provided enough information to calculate direct hits. A final homogeneous database with 29 studies was used (see Table 1.1) and also used in further analysis, including the comparison of the ganzfeld with other experimental conditions, selected vs. unselected participants and testing for experimenter effects (Baptista et al., 2015; Storm et al., 2010b). Looking at experimental conditions, Storm and colleagues compared the mean effect sizes of different experimental conditions and found that the ganzfeld had the highest *ES* (0.142), followed by non-ganzfeld noise reduction (0.110) and standard free-response (-0.029). However, only the difference between the ganzfeld and standard free-response, which do not feature noise reduction, was significant (ES mean difference = 0.17, $p < .001$). Nonetheless, these findings confirm that the ganzfeld is still the best developed method of free-response designs and that sensory isolation elicits increased psychic functioning (Baptista et al., 2015; Storm et al., 2010b).

For the analysis on experimenter effects, the authors divided the ganzfeld and non-ganzfeld studies into seven mutually exclusive groups (separated by

experimenter/laboratory), with at least two studies in each. The authors found no significant differences between the effect sizes across the groups, suggesting that there was no major effect of experimenters on study outcomes in the ganzfeld (Storm et al., 2010b). The authors also combined their 29-study database with the Storm and Ertel (2001) 79-study database to create a heterogeneous database of 108 ganzfeld studies. This database was used to test for decline effects across the ganzfeld from 1974 to 2008: using linear regression, they concluded that there was a significant correlation between year of study and effect size ($r(106) = -.21$, $p < .05$, two-tailed). This indicated a linear decline in effect size over a 34-year period. However, the authors also commentated on a rebound effect in recent years. When removing outliers, the linear correlation became non-significant, albeit still negative (Storm et al., 2010b). This rebound effect can be explained by the significant rise in effect size between the Milton and Wiseman (1999) and Storm et al. (2010b) databases, $r = .27$, $p < .05$ (Baptista et al., 2015). Thus, these findings suggest that the ganzfeld has not been strongly affected by a decline and remains one of the most consistent and reliable paradigms in parapsychology (Baptista et al., 2015; Storm et al., 2010b).

Several years later, Baptista and Derakhshani (2014) explored Storm, Tressoldi and Di Risio's 1997-2008 database, mainly questioning the validity of their sceptical hypotheses. Looking at the decline effects, Baptista and Derakhshani (2014) plotted study *ES*s against study publication year and found no decline in effect sizes ($r = .00$), contrary to Storm et al. (2010b). However, this database was found to be significantly heterogeneous and discovered this database produced two homogeneous subgroups - selected and unselected participants. Analysis on these subgroups found that the mean quality rating of studies with selected participants (weighted by sample size) was not lower than the mean quality rating for unselected participants. With a sceptical hypothesis one would expect *ES*s to decrease across the years as quality ratings increased: a small, non-significant correlation was found between ES and study year with selected participant studies ($r(12) = -.30$, $p = .29$). However, a non-significant positive correlation was found between study quality and ES ($r(12) = .27$, $p = .37$) and a similar correlation between study quality and study ES $r(12) = .26$, $p = .37$. The authors concluded there was no reliable findings but recommended more studies with selected participants are needed. Nonetheless, they concluded with "high confidence" that the sceptical hypothesis involving a decline in effect size and relationships between study quality and ES was not supported by Storm and colleagues database (Baptista & Derakhshani, 2014).

More recently, Cardeña (2018) analysed a database, containing 102 studies from the combined Storm and Ertel (2001) and Storm et al. (2010b) databases, was found to be highly significant in favour of the psi hypothesis. Cardeña (2018) confirmed the findings of Storm and colleagues when he analysed their database. Thus, the findings from studies using different protocols and researchers provides cumulative support for psi (Cardeña, 2018). Furthermore, selected participants had a bigger effect size, *ES* = 0.26, than their unselected counterparts, *ES* = 0.05, $t(27) = 3.44$, $p = .002$, confirming Baptista et al. (2015)'s re-analysis of the Storm et al. (2010b) database. The most recent meta-analysis by Storm and Tressoldi (2020) updates their earlier meta-analysis (Storm et al., 2010b). Once again, the authors found that the ganzfeld studies which recruited participants on selected characteristics produced a stronger mean effect than unselected participants.

Table 1.1

Summary of Meta-Analyses of Ganzfeld Research after the Joint Communiqué

Database	Years	Studies	Hit rate (%)	*ES*	Stouffer's Z	*p*
Bem & Honorton (1994)	1983-89	10	32.00	0.16	2.89	0.002
Milton & Wiseman (1999)	1987-97	30	27.60	0.013	0.70	0.24
Milton (1999)[1]	1987-97	38		0.25	1.45	0.074
Bem et al. (2001)	1987-97	40	30.10	0.051	2.59	0.0048
Bem et al. (2001) new	1997-99	10	36.70	0.17	3.97	<.001
Storm & Ertel (2001)	1982-1997	79		0.138	5.66	<.001
Storm et al. (2010b)[2]	1997-2008	29	32.2	0.142	5.48	<.001
Cardeña (2018)[3]	1997-2008	102		0.135	8.13	<.001
Storm & Tressoldi (2020)	2009-2018	9	31.00	0.119	2.56	<.001

Note. [1] Excluding the Dalton 1997 study. [2,3] Homogeneous database.

1.2.4 Conclusions and implications of Ganzfeld research

Overall, the described findings "provide cumulative support for the reality of psi" (Cardeña, 2018, p. 663). The evidence for psi is comparable with that for established phenomena in mainstream psychology, such as social psychology experiments which have an average effect size of 0.21 and other disciplines (Richard et al., 2003; Spencer, 1995). Furthermore, the psi ganzfeld data shows that with increased study quality *ES* improves, effect sizes have increased on average in recent decades, psi does not decline over the course of a long study and there is no file-drawer problem (Baptista et al., 2015).

Two other main conclusions can be drawn from these databases. First, there are discrepancies between certain databases. Honorton (1985)'s meta-analysis concluded "steps toward replicability of psi effects" (Storm & Ertel, 2001, p. 424), whereas others have found no replicable evidence (Milton & Wiseman, 1999). Later articles claim that these null findings were the result of including non-standard studies (e.g., studies using musical targets; Bem, 2001). This may reflect wider issues pertaining to meta-analyses and subjectivity in selection criteria (Murray, 2011). Research fields with greater flexibility in design, definitions and outcomes, the less likely the findings are to be true (Ioannidis, 2005). For example, the Bem et al. (2001) analysis reported the degree to which a replication adhered to the 'standard ganzfeld protocol' was positively and significantly related to effect size. However, the rationale for their investigation to study standardness was based on an internet debate between parapsychologists - this debate highlighted that researchers within the field would fail to agree on a single definition of the standard ganzfeld procedure (Milton, 1999; Schmeidler & Edge, 1999). Further, Bem et al. (2001) did not create their own definition for the raters. Instead, the authors provided the judges an extract from Bem and Honorton (1994)'s paper, with a different set of instructions when judging the PRL database, such as regarding creative participants as standard even though only two out of the nine series used a creative population. Equally, the addition or exclusion of certain studies (e.g., Dalton, 1997) can have a substantial effect on the cumulative findings (see Milton, 1999). The presence of greater flexibility increases the likelihood for transforming negative findings in to positive ones. Thus, adherence to common practices and standards is likely to increase the proportion of true findings (Ioannidis, 2005).

Alternatively, these differences between databases may be capturing questionable research practices (QRPs), as meta-analyses are sensitive to the accumulation of small,

systematic errors (Bierman et al., 2016). Bierman and colleagues simulated the presence of QRPs in ganzfeld telepathy research, simulating practices that had been reported in mainstream psychology (John et al., 2012). These practices include confirmation studies optimally stopped and turned into pilots, pilot studies that are included in confirmatory analysis, optional stopping, optional extension, publication bias, excluding data post-hoc and fraud. In the simulation of QRPs, each trial in a simulated experiment had the probability of a hit pre-set to 25% when simulating no anomalous ganzfeld effect or unknown QRP. For each QRP, there was a prevalence figure representing the probability an experiment would be 'using' this particular QRP. In each simulated experiment, a random decision was taken as to whether to apply each QRP, with the probability of each QRP equal to that of their prevalence in general psychology. When these practices were applied, the fit of the experimental database with the simulated database increased, apart from one practice, as shown in Table 1.2. The authors used the Storm et al. (2010b) experimental database, excluding studies pre-1985 due to their methodological weaknesses.

Table 1.2

Influence of QRPs on Simulated Telepathy Experimental Data

QRP	Relative gain in fit	Increase in Hit Rate over 25% (MCE)
Confirmatory to pilot	+73%	+2.8%
Pilot to confirmatory	+30%	+0.8%
Optional stopping	+35%	+0.9%
Optional extension	-63%	-0.3%
Publication bias	+88%	+3.5%
Exclude data post-hoc	+51%	+1.2%

Note. Adapted from Bierman et al. (2016), p.12. CC BY 4.0. Relative gain in fit calculated as fit = (fit with QRP - fit with no QRP)/fit with no QRP.

The authors concluded that their simulations suggested an unexplained excess in the hit rate of 2% and the simulated QRPs are capable of explaining 60% of the effect size reported in the ganzfeld meta-analysis. Thus, the highly significant probabilities reported in the ganzfeld meta-analyses are likely inflated by QRPs, although the overall result remains significant (p = .003) when including these practices. If true effect sizes are very small within a scientific field, it is likely this field is plagued by pervasive false positives (Ioannidis, 2005). Therefore, QRPs may account for large fractions of small effect size phenomena and

practices, such as confirmation to pilot and pilot to confirmation being the most likely culprits behind small effect sizes observed in mainstream psychology, even with large samples (Bierman et al., 2016).

However, Bierman and colleagues' simulation is not without criticism. Palmer (2016) argues that the Bierman paper is accusing experimenters of fraud, without evidence, and were making an effort to add QRPs to their models. Further, there are inherent issues with the notion of 'questionable research practices' - the authors' proportions of QRPs in their models were influenced by their judgements of whether QRPs were committed in the individual experiments. If Bierman and colleagues were truly interested in the QRP, they would have examined each experiment method section, contacted the authors if they were unsure about the procedure and decide how defensible their actions were and, removed the QRP label from that experiment. If these steps had been taken, this would have reduced the number of studies identified as containing QRPs, adjusted their model parameters thus, reducing the chance of explaining away the significance in the ganzfeld database (Palmer, 2016). A similar critique is offered by Bancel (2018), who conducted simulations of the QRPs proposed by Bierman and colleagues (2016). Operating under an extreme scenario in which QRPs were occurring at 100% prevalence, Bancel found that the simulation provided a poor fit to the ganzfeld data. He concludes that the simulation and QRPs outlined by Bierman do not offer a sufficient explanation for the ganzfeld findings and so, strengthen the case for the existence of psi using the well-established ganzfeld paradigm.

Furthermore, Palmer regards that only two of the QRPs can be truly regarded as fraud. The QRP 'fraud' is explicitly labelled and optional stopping can too be regarded as fraud, as many experimenters know that optional stopping can bias data. However, it is important to note that what makes a QRP fraudulent is its intent, not its nature (Palmer, 2016). If an experimenter is keeping record of the study and stops because they know it would produce a highly significant outcome, this implies intent, thus fraud. However, if the researcher keeps themself blind to the study record and has a legitimate reason to stop the study, then this too would be labelled as a QRP by Bierman. Although Bierman et al. (2016) acknowledge that the blind removal of data before inspection would not introduce bias, they later state that ganzfeld researchers are "generally not blind to the outcome of a session" (p. 19). And so, Palmer argues that "fraud must be detected, not inferred" (p. 12) and requires a case-by-case analysis into suspect researcher behaviour, rather than attempting to reliably conclude the existence of fraud from inference, in the absence of evidence.

One suggestion to minimise the culmination of small systematic errors in meta-analyses is to assess individual study power before inclusion (Muncer et al., 2002). Muncer et al. (2002) recommend to relax power requirements into a meta-analysis to 0.50. Applying this rule to the ganzfeld database, Bierman and colleagues found only six studies, contributing 748 trials, qualified and had a hit rate of 31.2% ($p < .001$), assuming no QRPs. When accounting for QRPs with the same prevalence as the simulated data, the hit rate dropped to 27.1% ($p = 0.07$). Thus, these simulations suggest a systematic problem within psychology and parapsychology. A survey of completed studies at the Koestler Parapsychology Unit at the University of Edinburgh revealed that 15% of non-significant studies were reported, compared with 70% of significant studies (Watt, 2006). Since this report there has been greater emphasis on the pre-registration of studies and detailed, open-source documentation of results (Watt & Kennedy, 2015; Wiseman et al., 2019) and greater methodological improvements within parapsychology, including defining study power and exploratory and confirmatory hypotheses (Kennedy & Watt, 2018).

1.3 Discussion

This chapter introduces the psi ganzfeld paradigm as the core experimental paradigm within parapsychology, but it is also an example to explore broader debates around scientific methodology, replication and evidence evaluation. Parapsychological research—particularly psi research—has long existed at the margins of mainstream science, often serving as the 'deviant science' (Schooler et al., 2018) and the researchers playing the role of 'academic jesters' (Wagenmakers et al., 2015). Research from parapsychology, namely the psi ganzfeld, is a unique example to explore decades of criticism, methodological reform, and epistemological challenges. The psi ganzfeld literature holds a unique illustrative position within parapsychological research and demonstrates the cyclical nature of the field and its internal dynamics.

As outlined in section 1.2.3, psi research follows a recognisable pattern: compelling evidence for psi is published, followed by critical evaluation, replication failures, methodological revisions, and eventual abandonment of a methodology – only for renewed interest to emerge after subsequent positive findings (Reber & Alcock, 2020). However, the psi ganzfeld stands apart from other short-lived methodologies, earning an "old faithful" status due to its continuous replication since its introduction in the 1970s. Even as newer experimental designs, such as Bem's (2011) "Feeling the Future" studies, which aimed to be

less resource and labour intensive, failed to replicate, researchers reliably returned to the ganzfeld paradigm, which has consistently produced statistically results (see Table 1.1). This not only demonstrates the psi ganzfeld's replication record but also its symbolic role within the discipline. During the replication crisis of the 2010s anomalous cognition studies published in 'respectable' journals were often framed as the sign of the breakdown of the scientific process (Schooler et al., 2018). Rabeyron (2020) highlights that psi researchers often face increasing expectations from the mainstream psychology (and scientific) community, such as more data, improved methodology, and more robust statistical approaches (e.g., Wagenmakers et al., 2011, 2015). Yet, as Schooler and colleagues (2018) suggest, such findings may represent a gradual accumulation of evidence that warrants continued exploration, rather than premature dismissal.

Interestingly, the parapsychology community also imposes high expectations on itself. McClenon (1986) observed that parapsychologists often engage in substantial self-critique, with a strong emphasis on methodological transparency and error detection. For example, Hyman's (1985) review concluded that there was no replicable evidence of the psi ganzfeld effect and findings were due to methodological weaknesses (now commonly referred to as QRPs). This sentiment was countered by Honorton (1985), who argued that the scientific method itself could resolve such disputes. Later, the null findings reported by Milton and Wiseman (1999), triggered a debate within the field about whether their analysis decision making inadvertently contributed to their results (see Milton, 1999; Wiseman, 2010).

Another unique attribute of psi ganzfeld research is the lack of scientific consensus over how to interpret its findings due to a lack of an accepted scientific model for psi and its source (Broughton, 1979; Palmer, 1997; Rabeyron, 2020; Reber & Alcock, 2020). 'Skeptics' often attribute significant results to bias, methodological flaws, or QRPs (e.g., Alcock, 2003; Hyman, 2010; Reber & Alcock, 2020), while 'proponents' argue that the consistency provides support for the existence of psi, and that future work should focus on the process of psi, not more evidence that it exists (e.g., Cardeña, 2018; Cardeña et al., 2014; Radin, 2006). There is also a third argument that parapsychologists have moved too quickly into empirical work without first developing a testable theory (Braude, 1992). Even more, parapsychologists should be 'good observers' and study human behaviour in real-life contexts rather than experimental paradigms which are only of significance to the experimenter (Braude, 1992). Regardless of these positions, Schooler et al. (2018) argues the implications of psi ganzfeld are valuable: either the results represent genuine phenomena that challenge current scientific

models of consciousness, or they indicate a longstanding pattern of systematic error conducted by researchers using scientific methods. In either case, the psi ganzfeld serves a valuable case study for examining replication, researcher degrees of freedom and experimental research within the broader landscape of psychology.

1.4 Thesis overview

The remainder of this thesis builds on the themes and argument introduced in Chapter 1 that the psi ganzfeld is a unique experimental paradigm with a long and public discourse surrounding published results. The psi ganzfeld is a case study which explicitly demonstrates a pattern of claims of anomalous phenomena in a controlled laboratory setting, with a proceeding backlash (both within and outwith the academic field), which insinuates questionable research practices and cherry picking.

Chapter 2 serves as an extended introduction, situating the ganzfeld paradigm within the broader context of the replication crisis. This chapter explores methodological and statistical challenges faced by researchers in both parapsychology and mainstream psychology, with emphasis on questionable research practices, researcher degrees of freedom and analytic flexibility.

Chapter 3 is the first empirical chapter and was initiated prior to the COVID-19 pandemic. The meta-analysis reported in this chapter (published as Pooley et al., 2023) was originally designed to inform a large-scale telepathy study. The analysis focuses on how design variability across previous telepathy studies influenced study outcomes and emphasises the need for more transparency from researchers. This meta-analysis found that a formal mentation review period decreases the likelihood of study success, whereas allowing the sender to listen to the participant's verbal report increases study success.

Chapter 4 builds on the concerns encountered during the analysis in Chapter 3 and issues raised in Chapter 2, particularly the extent to which analytic flexibility affects meta-analyses. This chapter presents a methodological umbrella review of all published psi ganzfeld meta-analyses, with a focus on the role of inclusion criteria and post-hoc decision making on influencing outcomes. This review found that the psi ganzfeld meta-analyses are heterogenous in the studies they include and are often of poor quality.

Chapter 5 focuses on an underreported area of experimental psychology: software validation. Building on the issues discussed in Chapter 2, the experimental software and data

from a large-scale precognition study conducted at the Koestler Parapsychology Unit (KPU) are used as a practical demonstration of how experimental software can be tested for integrity. This testing involves ruling out subtle biases in the experimental software, a practice not routinely conducted in psychology experiments.

Chapter 6 introduces a specification curve analysis to further investigate the influence of researcher degrees of freedom, even in preregistered studies. Using psi ganzfeld data, this chapter illustrates how different analytic choices can provide varied results, thereby reinforcing concerns about selective reporting and flexibility in statistical modelling.

Chapter 7 presents a novel quantitative language analysis of verbal reports (mentations). Given the lack of a theoretical model of psi, this chapter uses quantitative language analysis to examine these verbal reports as potential indicators of underlying cognitive or anomalous processes, contributing a new perspective to the interpretation of ganzfeld results. This analysis found that psi ganzfeld participants are on average, anxious but honest in their verbal description of the ganzfeld stimulation and there is no clear differentiation between those who are successful in the task than those who are not.

Finally, Chapter 8 synthesises the findings of all empirical chapters, offering an integrated discussion of how the psi ganzfeld paradigm exemplifies broader concerns in psychology research, and reflects on the implications for future research.

Chapter 2

Addressing Questionable Research Practices (QRPs)

2.1 Chapter overview

As outlined in Chapter 1, questionable research practices (QRPs) are not issues restricted to parapsychological research: social psychology, cognitive neuroscience and medical research are too plagued by these practices (Bierman et al., 2016). This chapter focuses on methodological and statistical weaknesses that blight both psychological and parapsychological research. Due to the unique nature of psi research, which may be trying to examine a phenomenon that does not exist, it routinely sparks debate in the wider psychological community. Common critiques are poor methodology and subpar statistical practices which may be producing inflated effects. This chapter details recommendations and potential solutions that were published in the wake of the 2010s replication crisis era and how these can be addressed, using parapsychology as an example.

2.2 Psychology's replication crisis

John et al. (2012) surveyed psychologists about their research behaviour. From approximately 2000 respondents, 1 in 10 research psychologists had introduced false data into a scientific record and an overwhelming majority had engaged in selective reporting of studies, collected more data after determining if the results were significant, and reported unexpected findings as having been pre-specified. Ninety-four percent of respondents in the Bayesian-truth-serum condition, a scoring algorithm which provided incentives for truth telling, admitted to having engaged in at least one QRP. When asked about their actions being defensible, respondents who admitted to a QRP tended to agree and these findings generally did not differ across discipline or type of research (John et al., 2012). As a knock-on effect, meta-analyses will undoubtedly reflect these inflated, and perhaps, falsified findings (Kvarven et al., 2020). Small studies usually have fewer experimental personnel and less formal procedures than larger, more sophisticated studies and are thus, more susceptible to some form of experimenter effect, including fraud (Ioannidis, 2005; Kennedy, 2013).

With such a high prevalence of QRPs and weak methodologies within psychological research, there is a greater incentive for pre-registered, well-powered confirmatory experiments, akin to medical research (Kennedy, 2016).

For a long time, psychologists have been aware of a paradox within published research. The overwhelming majority of published findings are statistically significant, whilst the vast majority of published studies are under-powered (Fanelli, 2012). Thus, it is theoretically unlikely to obtain results that are statistically significant (Chase & Chase, 1976). *P*-hacking (the selective reporting of data and analyses) is the most likely answer to this paradox: it is the only practical way that would consistently produce statistically significant findings from under-powered studies (Nelson et al., 2018; Simmons et al., 2011).

Furthermore, a number of authors (Bierman et al., 2016; John et al., 2012; Wagenmakers et al., 2011) have discussed the notion that psychological research adopts data-driven analyses. These practices include noticing apparent patterns in the data and testing them for significance, the inclusion or exclusion of variables amongst others. Some of these actions may be warranted depending on the circumstances, but even in the best of cases, they complicate matters. These practices make it harder to assess the accuracy of an individual study as they generally increase the probability of obtaining a significant finding (Maxwell et al., 2015). *P*-hacking may be an attractive option for some researchers: significant results may secure funding or open the door to a promotion. All the while, leading funding bodies to direct their resources away from hypotheses that are actually true and policy makers implementing potentially harmful or ineffective policies (Nelson et al., 2018; Simmons et al., 2011).

Attempts at replicating past study findings more often than not fail to show the same results, leading to questions about the credibility of the original results (Maxwell et al., 2015). For example, a large collaborative project in 2015 was conducted to assess the reproducibility of psychology studies conducted in three high-ranking psychology journals in 2008 (Open Science Collaboration, 2015). The researchers conducted replications of 100 experimental and correlation studies, using high-powered designs and original materials, where available. In the original studies, 97% were reported as significant, whereas only 36% of the replications had statistically significant results with effect sizes half the magnitude of the original effects.

Low-power research designs, paired with publication bias favouring positive results, create a literature with upwardly biased effect sizes. Thus, replication effect sizes are bound to be smaller than the original studies. This is not due to differences in implementation, but due to the original study reporting inflated *ES*s driven by publication bias, whereas the replications are not (Open Science Collaboration, 2015). The high rate of non-replication in research is a consequence of claiming conclusive findings solely on a significant *p*-value, usually less than 0.05 (Ioannidis, 2005).

2.2.1 Meta-analyses

This replication crisis is exacerbated by meta-analyses (Sharpe & Poets, 2020). Meta-analyses (MA) allows researchers to assess methodologically similar studies and review the collective findings, thus they are a form of *post-hoc* analysis (Kennedy, 2013). As with other types of *post-hoc* analysis, MAs require the researcher to make a multitude of decisions from methodological decisions, study selection criteria, statistical methods and more. Most decisions do not have apparent right or wrong answers and so, different choices can lead to different outcomes, resulting in ambiguity and a potential for making biased decisions (Kennedy, 2013; Watt & Kennedy, 2016). One could argue that if individual studies in a meta-analysis are pre-registered then there would be little bias. However, this does not eliminate the decision making process of the research conducting the meta-analysis: they may be more likely to include the studies that report significant findings (Kennedy, 2013; Watt & Kennedy, 2016).

One potential solution is multiverse analysis. Data processing often involves making choices among several options from the exclusion, transformation and coding of data (Simonsohn et al., 2020; Steegen et al., 2016; Voracek et al., 2019). Thus, instead of performing a singular analysis, one could perform a multiverse analysis whereby the researcher conducts analysis across the whole set of alternative data sets that correspond to a large set of reasonable scenarios. Alternatively, one could perform a prospective meta-analysis. A prospective MA is a form of preregistered confirmatory research as the data and statistical analysis plan is pre-specified and publicly registered before the included studies are conducted (Watt & Kennedy, 2016). Inclusion into the prospective MA is already determined by a set of inclusion criterion before the results of the studies are known, reducing bias that may occur during the study inclusion process.

An example of the issues with retrospective meta-analyses is a study which compared the effect sizes of large-scale registered replication studies in psychology with a meta-analysis of these replications (Kvarven et al., 2020). The authors found that the reported mean MA effect size was almost three times larger than the mean *ES* of the replication studies. And so, meta-analyses are ineffective at fully adjusting for inflated effect sizes caused by publication bias and selective reporting (Kvarven et al., 2020).

Meta-analyses, which were dubbed to be "controversy killer[s]" (Broughton, 1991), have not been effective at resolving scientific controversies, thus retrospective meta-analyses cannot be a substitute for well-powered confirmatory experiments (Ioannidis, 2005; Kennedy, 2013). Ioannidis stated that there is a major problem in scientific research as it is impossible to know with 100% certainty what the truth is in any research question, thus the 'gold standard' is unattainable. Nonetheless, to try and improve the situation of low replication success, better powered evidence is a must. Large studies, or low-bias MAs, will help bring us closer to the unknown 'gold standard'. Additionally, reducing bias via enhanced research standards and curtailing of prejudices are also beneficial (Ioannidis, 2005; Panagiotou & Ioannidis, 2012).

2.2.2 Replication in parapsychology

As shown in chapter 1, various conclusions have been made about replication within the ganzfeld literature. Whilst some researchers conclude there is little evidence of replication (Milton, 1997b, 1999), others declare there is substantial evidence of psi in the ganzfeld (Cardeña, 2018). Sceptical psychological researchers will advocate for the improvements of methodological practices, rather than assigning experimental results to psi. However, this inefficient cycle of improving minor methodological practices leads nowhere. Instead, there are larger methodological factors both psychologists and parapsychologists have to face that are standard procedure in medical research (Kennedy, 2016).

One major feature of regulated medical research is the pre-registration of studies. Put simply, pre-registration is the detailed specification of one's research plans in advance of data collection and submitting the plans to a registry (Center for Open Science, n.d.). By pre-registering a study, one is specifying exploratory (hypothesis generating) and confirmatory (hypothesis testing) research - the same data cannot be used to create and test a hypothesis (Center for Open Science, n.d.). Formal study registries for psychological research include

the Center for Open Science which went live in 2013, and AsPredicted. For parapsychological research, the Koestler Parapsychology Unit's Registry began in 2012 and welcomes all forms of parapsychological research and the *Society for Psychical Research's* data repository, Psi Open Data, providing an open repository for parapsychological and psychical research data.

In response to the replication crisis, research bodies within psychological sciences responded differently to pre-registration requirements. The response to QRPs varies between journals, for example *Psychological Science* encourages the use of effect sizes, estimation, meta-analysis and extra detail of study methods (Giofrè et al., 2017). Whereas, some journal guidelines may refer to a third-party, such as American Psychological Association (APA)'s publication manual (Giofrè et al., 2017). Registered reports (RR) are more journal-specific, whereby a researcher's study protocol is subject to peer review prior to data collection. Studies may be provisionally accepted to be published, based on the quality of the submitted protocol (Nosek & Lakens, 2014).

Wiseman et al. (2019) compared studies over a seventeen-year period within the *European Journal of Parapsychology* that were registered reports and those that were not. With a total of 60 studies (25 registered reports and 35 non-registered), the authors found that 28.4% of the statistical tests in the non-RR group were significant (95% CI [21.5%–36.4%]); compared to 8.4% of those in the RRs (95% CI [4.0%– 16.8%]). There was no effect due to studies researching different topics or improved methodology over time. Thus, these findings are consistent with the concept that pre-registration (or registered reports) reduce questionable research practices. Furthermore, 8.4% of studies in the RR group were significant, which is greater than the chance expectation of 5%, suggesting that pre-registered studies produce higher quality evidence than non-RR studies (Wiseman et al., 2019). Therefore, it is clear that questionable research practices are inflating results and meta-analyses are failing to adjust for this. Retrospective meta-analyses cannot substitute for well-designed and powerful individual studies (Kennedy, 2013).

As identified by Wagenmakers et al. (2011), and if Bierman et al. (2016) simulations are accurate, their simulated analyses revealed the amount of bias captured in the ganzfeld literature, inflating the reported findings. And so, researchers must attempt to quell these biases by addressing statistical, methodological and security issues that many psychological researchers have failed to acknowledge (Kennedy, 2014b; Watt & Kennedy, 2016).

Retrospective meta-analyses have failed to resolve scientific controversies as shown in the psi ganzfeld database: one may obtain a significant result by combining multiple studies, but these are likely to have captured experimenter biases and excessive exploration that exists within these experiments (Wagenmakers et al., 2011; Watt & Kennedy, 2016). Hence, researchers must preregister their reports (Watt, 2005; Wiseman et al., 2019), ensure they have optimal power (Muncer et al., 2002) and declare their analysis plans before data collection to help reduce the accumulation of small systematic errors.

2.3 Methodological and statistical issues in psi research

2.3.1 Methodological weaknesses

A common criticism of psi research is that a replicable effect has not been demonstrated and there are errors in the experimental design (Bem & Honorton, 1994; Milton, 1999). Methodological weaknesses identified in psi research include sensory leakage, poor target randomisation procedures and experimenter fraud.

2.3.2 Experimenter fraud

Having previously highlighted the abundance of QRPs in mainstream psychology (Bierman et al., 2016; Milton, 1999), it is no surprise that experimental research in parapsychology has certain characteristics that could be labelled as fraud or misconduct, especially as a fraudulent researcher could claim psi is responsible for the findings (Kennedy, 2014b). Experimenter fraud is an acknowledged factor in scientific research (Stroebe et al., 2012), however the full extent of fraud is unknown as undetected instances are likely (Kennedy, 2014b). Standard research procedures in psychology and parapsychology do not include measures to prevent and/or detect experimenter fraud, thus offering an easy and potentially tempting opportunity for fraudulent behaviour with little possibility of detection (Kennedy, 2016). Independent replication and peer review have failed at detecting and averting experimenter fraud (Bierman et al., 2016; Open Science Collaboration, 2015; Wagenmakers et al., 2011).

Most parapsychological experiments have not obtained statistically significant results, however a few experimenters have reported significant results on almost every experiment (Kennedy, 2013). One such example is the case of W. J. Levy, who was obtaining significant results on every experiment (Kennedy, 2014b). The Levy case was different to usual, *post-*

hoc fraud investigations as direct evidence of fraud was obtained whilst the fraud was occurring, such as the tampering of laboratory equipment. Further, suspicions were raised about Levy's analysis software - Kennedy suspected the software (designed by Levy) was developed to produce fraudulent results, an easily detectable offence. It turned out that Levy was in fact fabricating published results, with no effort to make them match with the logged data or analysis program data copy (Kennedy, 2014b). There have been other cases of fraud published within parapsychology (Rhine, 1974, 1975).

Evidence of fraud from *post-hoc* investigations are likely to be unconvincing in parapsychology, as a fraudulent researcher could claim psi was at work, rather than fraud (Kennedy, 2014b). On the other hand, declaring suspicions of fraud without substantial evidence can create an intolerable environment which leads to everyone involved being discredited (Kennedy, 2014b). And so, some argue that you cannot make firm conclusions about the existence of fraud from inference in the absence of evidence, such as the simulation of QRPs by Bierman et al. (Palmer, 2016). Thus, fraudulent behaviour must be determined on a case-by-case basis by reviewers checking for possible fraudulent behaviour in published reports and following up with the author, and by lab staff who may notice irregularities in data and following up on their observations (Palmer, 2016).

Kennedy (2014a) discusses two categories of experimenter misconduct, *data manipulation* and *analysis manipulation*. Data manipulation is overt fraud conducted by an experimenter(s). The extent of fraud is difficult to quantify, undetected instances are probable and cases of recognised fraud are unlikely to be published. Fraud is more likely to occur if the risks of detection are low and when there are financial incentives or pressures to produce a certain result (John et al., 2012; Stroebe et al., 2012). Independent replication and peer review are not deterrents to fraud, as demonstrated by the inflated effect sizes reported within psychology and parapsychology (Bierman et al., 2016; Kvarven et al., 2020; Wagenmakers et al., 2011). Analysis manipulation covers a range of practices which may significantly distort reported findings. Although analysis manipulation is inexcusable and can be easily prevented (Kennedy, 2013), it is still a widespread occurrence in psychology (John et al., 2012). Practices include:

- Analysing multiple different hypotheses and only reporting those that support the experimenter's expectations (without mentioning or correcting for other analyses)

- Failing to report experiments with results that fail to support the experimenter's expectations
- Planning a vague hypothesis and then determining the specific statistical test and hypothesis whilst exploring the data during analyses
- Reporting exploratory or *post-hoc* findings in a way that can be mistaken for planned analyses
- Adapting the description of the methodology or findings to conform to referee comments during the publication process

Study preregistration is a simple and effective way of deterring analysis manipulation however, in wider psychological research there is little mainstream use of preregistration (Kennedy, 2013). In a survey of 1035 social and personality psychology researchers, only 27% of researchers had preregistered their study (Washburn et al., 2018). When asked about their rationale for not preregistering, 21% responded that preregistration of studies is not required and does not increase validity (i.e., there is no incentive to register and it is not commonplace to do so). Thus, there remains some hesitation and misconceptions about the preregistration of studies within psychology.

Another area of experimenter fraud that is often not reported, or even considered, is fraud within the experimental software. The documentation of software validation is vital to ensure confidence in published findings. This practice is customary in clinical research (U.S Food and Drug Administration, 2002) but is less recognised in psychological and parapsychological research. Much like assessing a questionnaire, software validation involves testing and documenting the software to ensure it behaves reliably, accurately fulfilling its purpose.

2.3.3 Sensory leakage

Sensory leakage is the transmission of information via normal communication methods, however a rigorous anomalous cognition experiment must ensure there is no communication between the receiver, the target, the sender and anyone who may have knowledge of the target, at all stages of the experiment (Holt et al., 2012). Sensory leakage is the most insidious of problems as it is nigh on possible to completely eliminate, as it may occur in subtle or unforeseen ways. The autoganzfeld studies of the mid-1980s to mid-1990s

were designed to reduce methodological and security problems of the early ganzfeld research, highlighted by Hyman and Honorton (1986). However, issues of sensory leakage, inadequate shielding and questionable research practices still lingered according to some critics (Bierman et al., 2016; Milton, 1999; Wiseman et al., 1994). For example, participants familiar with the laboratory layout could communicate via structural connections to send messages between the sender and receiver. Even if the receiver is adequately shielded, a low-level sound may act as a cue to unconsciously influence the receiver during the judging stage (Wiseman et al., 1994).

By automating the procedure, such as target randomisation, selection, presentation, judging, and responses recorded by the computer software, it allows the sender to remain isolated once the session has commenced. The target is withheld from both the experimenter and receiver until the end of the trial (Holt et al., 2012). However, there may be as yet unidentified sensory leakage with modern automated ganzfeld procedures. For example, a telepathy study using systems that transfer data, such as the receiver's mentation to the sender, could be interfered with (intentionally or not) and thus, jeopardising the secrecy of the target.

2.3.4 Randomisation

Anomalous cognition studies statistically compare a series of guesses to chance expectation, thus sufficient randomisation of targets is essential (Holt et al., 2012). Systematic patterns in target selection must be prevented to ensure they coincide with response biases or are detectable by the receiver. This is less of an issue in ganzfeld studies as there is only one target per trial, however if the study uses the same sender or independent judge for all trials, then there is a risk the non-random target selection will coincide with the non-random person's choice (Brugger & Taylor, 2003; Holt et al., 2012).

Selection of a target must also be independent of any predispositions of the participants within the ganzfeld study. For example, if the sender and receiver were colluding and the sender was allowed to select their own target, the receiver may be able to discern the target due to a shared taste (Holt et al., 2012; Milton, 1997b). There are a variety of methods to produce random sequences, however they all have unique flaws. For example, electronic random number generators can develop intermittent faults, have design flaws and be susceptible to environmental influences (Milton, 1997b).

2.4 Potential solutions for methodological weaknesses

The identified methodological weaknesses have resulted in a wide range of potential solutions and recommendations. The flaws mentioned previously are not definitive, however they pose some of the greatest risks to a psi ganzfeld experiment. To address these issues, potential solutions will be discussed, namely from Kennedy (2016) and Milton and Wiseman's (1997) book *Guidelines for Extrasensory Perception Research*.

2.4.1 Deterring experimenter fraud

Given the controversial nature of anomalous cognition research, experimenter fraud is a major consideration when designing a ganzfeld study. Paramount to any AC study is the implementation of practices that prevent and deter overt experimenter fraud. Experimenter fraud should not be tempting or easy in a parapsychological experiment. The use of multiple experimenters is recommended in psi research as it reduces the likelihood of tampering and is a common practice in regulated medical research (Dalton et al., 1996; Kennedy, 2017). For psi experiments, it is easy to establish a procedure with multiple experimenters that:

1. Have duplicate copies of the randomisation and results
2. Check or observe each another
3. And/or switch roles in the experiment

Practices such as duplicate records and experimenters checking each other are easy to implement and help make undetected experimenter fraud harder, rather than easy and tempting (Kennedy, 2014b, 2016). Duplicate copies of each component of the data sent to a secure location as early as possible in the data collection is one such strategy. For example, data can be stored to an external server as well as locally, making copies of the critical data in two locations, and thus less susceptible to undetectable manipulation by an individual (Watt et al., 2020).

In addition to electronic duplicates, physical records logged by the experimenters (e.g., receiver's ratings, sender noting down the target number to ensure it matches with computer log) can provide further security by comparing for discrepancies before data analysis, or in case of failures in the electronic recording systems (Dalton et al., 1996; Watt et al., 2020).

Making the raw data available to other for independent analysis is useful too, however it is a secondary strategy to deter fraud. Making the data available does not eliminate the need for preventative measures.

2.4.2 Study Preregistration

One way to eliminate the issue of analysis manipulation is study preregistration (Kennedy, 2013; Wiseman et al., 2019). The prospective public registration of experimental studies is common within medical research, prior to first patient enrolment, as consideration for publication (International Committee of Medical Journal, 2020; Kaplan & Irvin, 2015). The merits of pre-registering confirmatory studies is widely recognised and allows confirmatory results to be given the full credit they deserve (Nelson et al., 2018).

Key features of a study registration includes a) the registry specifies requirements for the registration information, b) all key methodological decisions that could affect the study outcome are registered, c) each registration is independently reviewed, d) study registration is made irreversibly public before data collection begins and, e) the registrations can easily be found and accessed online. One such study registry is at the Koestler Parapsychology Unit (Watt & Kennedy, 2015). Registration of studies is important for both sceptics and proponents: sceptics sometimes use analysis manipulation to negate positive results (Kennedy, 2013; Milton, 1999). As previously shown, the preregistration of study helps alleviate systematic errors within parapsychological research (Wiseman et al., 2019). In parapsychological research it is now commonplace for studies to be preregistered: not doing so will likely generate scrutiny from the research community (Kennedy, 2014b).

2.4.3 Software validation

End-user testing is the most vital step to validate software whereby a tester, who did not do the software programming, assesses the software (Kennedy, 2016). End-user testing ensures the software operates as intended in the environment that will be used for the study. Testing usually reveals problems that the developer(s) did not anticipate and allows for necessary adjustments to be made before data collection. Software validation for automated experiments should assess the following criteria:

1. Does the software accurately and reliably present the stimuli and/or the feedback for the experiment?

2. Does the software properly generate the random elements in the experiment?

3. Does the software accurately and reliably record the human inputs and the conditions generated by the software?

4. Does the software properly handle unexpected, inappropriate inputs?

5. Does the software have these properties for all computers that will be used in the experiment?

End-user testing should detect intentional programming errors (fraud) as well as unintentional errors. Ideally, the software developer should not have access to the computers used for data collection. Likewise, the experimenters collecting data should not have access to software source code (Kennedy, 2016).

2.4.4 Sensory leakage

To protect the target and any collusion between researchers and participants, the receiver should be supervised by an experimenter during the trial and a report of who was present during each trial (Kennedy, 2016; Milton & Wiseman, 1997). When shielding the sender from the receiver, the sender must be instructed to be as quiet and still as possible, sender is supervised in a such a way that and obvious attempts to communicate with the receiver is detected and the sender and receiver must be in different rooms with at least one other room between them (Milton, 1997b). Visual and auditory shielding and supervision together should be sufficient to prevent an accomplice from detecting signals from the receiver. Further measures against cheating, such as using a Faraday cage to prevent radio communications and using a laboratory personnel as senders, can also be incorporated (Milton & Wiseman, 1997).

The final report should provide sufficient detail for readers to assess the quality of the procedures and preventative measures put in place (Kennedy, 2016). This includes details of room set-ups and locations to ascertain if sound or vibrations could be carried between the

sender's and receiver's rooms, measures to shield the sender and/or receiver, how the target was kept safe and full detail of the experimental procedure (Milton & Wiseman, 1997).

2.4.5 Methods of randomisation

Ganzfeld research using electronic random number generators (RNGs) must detail the model name and the computer interfaced with it, so the readers can assess the likelihood of design problems and access fuller documentation (Hyman & Honorton, 1986; Milton & Wiseman, 1997). In addition, extensive randomness checks should be performed prior to study commencement and at regular intervals thereafter (Milton & Wiseman, 1997).

2.5 Statistical issues

As shown, low-powered studies combined with publication bias has resulted in a psychology database with inflated effect sizes (Kvarven et al., 2020; Nelson et al., 2018). This is also evident in the psi ganzfeld database. Notable statistical issues in psi ganzfeld research include inadequately powered studies, over-reliance on null hypothesis statistical testing (NHST) and the blurring between exploratory and confirmatory hypotheses.

2.5.1 Over reliance on NHST

The fundamental weakness of NHST is that a failure to reject a null hypothesis does not establish that there is no effect (Carver, 1978). Likewise, a decision in favour of the alternative hypothesis is not necessarily enough to provide strong evidence that the alternative hypothesis is true and the null hypothesis is false (Neath, 2010). NHST can provide highly misleading evidence against the null hypothesis, leading one to reject the null hypothesis when there is no evidence to do so (Wagenmakers & Grünwald, 2006). The eagerness of the psi community to find any effect, regardless of how small, interacts with the weakness of null hypothesis significant testing (Wilson & Shadish, 2006).

The high rates of non-replication are a consequence of this convenient, but ill-founded method of claiming conclusive research finding purely on the basis of a single study assessed by formal statistical significance, usually a *p*-value below 0.05 (Ioannidis, 2005). Traditional *p*-values tend to overstate evidence against the null hypothesis and is exacerbated as the sample size increases (Wagenmakers et al., 2011).

2.5.2 Low powered studies

Within the parapsychology literature there are assumptions that in order to establish itself, the field needs to find a replicable experiment. However, it has never been clear what exactly would constitute as accepted evidence of a replicable experiment (Utts, 1991). From a statistical point of view, the problem remains of how to create a sufficient statistical power to identify or 'capture' a phenomenon (Tressoldi, 2016). Statistical power is the power to detect a given phenomenon after defining the probability of risk accepting the hypothesis that it is true, given an effect size when it is not (Tressoldi, 2016; Tressoldi & Utts, 2015).

Power analysis should be conducted prospectively, to calculate the minimum sample size needed so that one can increase the likelihood to detect an effect of a given size (Tressoldi, 2016; Tressoldi & Utts, 2015). Power depends on three factors:

1. The probability of a Type I error - the probability of accepting the existence of a phenomenon when is it is not, usually 5%
2. The size of the sample(s) used
3. An effect size parameter indexing the actual degree of deviation from the probability of 'non-existence' in the general population

The psi ganzfeld literature has repeatedly shown effect sizes of below 0.30, with the most recent update showing a combined homogeneous database *ES* of 0.135 (Cardeña, 2018). Even studies that only use selected participants have an effect size of 0.26 (Cardeña, 2018). Relating to the previous weakness, NHST does not provide an estimate of the difference from the null hypothesis (a measure of *ES*), even if there is a difference from zero. A *p*-value less than .05 does not mean a large effect, even if the *p*-value is very small (Tressoldi & Utts, 2015). This is a consequence of the *p*-value depending on the sample size for all situations, apart from when the null is exactly true (Tressoldi & Utts, 2015).

The smaller the *ES* in a scientific field, the less likely the findings are to be true, as power is related to *ES* (Ioannidis, 2005). Even if these small effect sizes are true, the more likely the field is plagued by ubiquitous false positives (Ioannidis, 2005). Likewise, the smaller the sample size of conducted studies, the less likely the findings are to be true as a smaller sample size equals smaller power. Hence, research findings are more likely to be true in fields that conduct studies with large sample sizes (Ioannidis, 2005).

2.5.3 Participant dropouts and incomplete data

Participant dropouts and incomplete data can introduce bias into experimental findings, however they are often given no attention by researchers. Dropouts and incomplete data cannot be assumed to be independent of the experimental condition, thus act as confounding factors that can add bias into the model (Kennedy, 2016). In clinical research the *intention-to-treat* principle states that once the participant is randomised into a condition, their data is analysed, whether they complete the study or not (U.S Food and Drug Administration, 1998). The inclusion or exclusion of incomplete data may alter the study outcome thus rendering the results as unconvincing. One caveat is that *intention-to-treat* is more likely to underestimate the analysed effect but generally more convincing that an effect has truly occurred.

Alternatively, *per-protocol* analysis is more commonly practised, where participant data is included if they adhered to the study protocol. If not, their data is removed. However, *per-protocol* has the tendency to overestimate the effect being researched by inflating effect sizes, as those with poor performance are more likely to drop out of a study.

2.5.4 Exploratory or confirmatory?

As reported by Bierman and colleagues (2016), the ganzfeld literature is host to a number of questionable research practices and their simulations found an unexplained excess in the hit rate of 2%. No single ganzfeld study has been conducted at this scale, given the financial constraints of parapsychology. However, combining a series of coordinated smaller studies as a replication effort could be statistically equivalent to a singular large study, only if QRPs like confirmation-to-pilot or pilot-to-confirmation are excluded (Bierman et al., 2016).

Exploratory analyses are useful as they allow for the creation and development of novel lines of inquiry. However, they usually involve methodological or theoretical ambiguities that can convolute the interpretation of findings (Koestler Parapsychology Unit, 2018). Exploratory analysis results presented as confirmatory disguises the fact that the researcher has analysed the data twice: first to discover a new hypothesis and second, to test the hypothesis (Wagenmakers et al., 2011). Hence, exploratory hypotheses should be explicitly mentioned and statistical results should be adjusted accordingly such that analyses are more conservative (Wagenmakers et al., 2011).

A well-designed confirmatory analysis should be able to provide evidence that an experimental hypothesis is false, as well as true (Koestler Parapsychology Unit, 2018). However, many studies in both psychological and parapsychological research are not well-designed confirmatory analyses as they often suffer with problems such as low power and uncertainty in the methods or theory (Kennedy, 2016; Koestler Parapsychology Unit, 2018). Thus, it is paramount that results are clearly indicated as exploratory or confirmatory.

2.5.5 Recommendations to prevent statistical flaws

Many of the following statistical recommendations are applicable to parapsychology and general psychology. As demonstrated by the critiques of parapsychological research (e.g., Bierman et al., 2016; Wagenmakers et al., 2011; Wiseman, 2010), psychology must change its statistical practices. Below are some recommendations for pressing issues, such as statistical practices for confirmatory research, performing Bayesian analysis and performing power analysis in advance of data collection.

Statistical methods for confirmatory research

In recent decades psychological researchers have excessively focused on *p*-values without reasonable consideration of effect size, power and distinguishing between exploratory and confirmatory research (Kennedy, 2016). As in regulated clinical research, the *p*-value, effect size and statistical power must all be considered and a clear distinction made between exploratory and confirmatory research (U.S Food and Drug Administration, 1998; Wagenmakers et al., 2011). As demonstrated by Bem's (2011)paper *Feeling the Future*, exploratory methods can provide a highly misleading image of study results by overstating the statistical evidence against the null hypothesis (Wagenmakers et al., 2011).

Many psychologists agree that falsifiable research is a basic goal of science, however the methodological and statistical practices to conduct falsifiable research have not been recognised and implemented by researchers (Kennedy & Watt, 2018). To conduct falsifiable confirmatory research, one must first select a minimum effect size of interest - such that any smaller effect would be too small of interest, or would provide evidence that the tested hypothesis is false. Second, a sample size that has power of at least 95% for the minimum effect of interest needs to be determined - failure to obtain a significant result with 95% power is reason that the expected effect specified in the power analysis is false under the

conditions of the study. An effect size with 95% power is the *falsifiable effect size* for a study. Operating characteristics, or a power curve, for a planned analysis will reveal the effect sizes that can be readily detected in a study - a step required both for Bayesian and classical statistics. Third, one should publicly preregister the study with specific inference criteria for evidence that the effect has or has not occurred (Kennedy & Watt, 2018).

When reporting statistical findings, it should be clear the initial choice of sample size and an explanation if it was not met and rationale for sample size, including power analysis. Further, confidence intervals should be reported as they provide the magnitude of effect and, the effect sizes along with or instead of *p*-values (Tressoldi & Utts, 2015). In addition, the data should be treated as *intention-to-treat* and at the very least, excluded data should be analysed and reported to evaluate possible bias. The key question here is to assess if the overall result is significant when the excluded data is included. For missing data, sensitivity analysis exploring various negative assumptions about the missing data should be conducted to evaluate any bias. The treatment of incomplete data should be specified in the study preregistration document and discussed in the final study (Kennedy, 2016).

Power and effect size

The criterion for successful confirmation of a hypothesis will be a significant result in a study with a power of 0.80, or greater. However, this recommendation was made in the 1960s when data collection and analyses were typically done by hand (Cohen, 1965; Kennedy & Watt, 2018). Likewise, the common practice of basing power analysis on the mean effect size from previous studies does not consider the uncertainty of the estimates or the likelihood of biased estimates due to QRPs (Kennedy & Watt, 2018; Yuan & Maxwell, 2005). As previously stated, a power closer to 95% or greater is required for falsifiable research (Kennedy & Watt, 2018). Well designed confirmatory research can provide evidence for and against a hypothesis (Watt & Kennedy, 2015). Interpreting non-significant study results with low power is ambiguous, as the results may be due to low power rather than the experimental hypothesis being false. *P*-hacking is the only practical way to consistently produce significant results from under-powered studies (Nelson et al., 2018).

For effect size, the *ES* used in power analysis when designing a study is a prediction about what will happen: a non-significant result in a study with high power provides evidence that the predicted *ES* specified in the power analysis is false (Kennedy, 2016). A non-significant result does not give evidence that the null hypothesis is true - a small, non-zero effect size

could be true. If a minimum *ES* is not specified for power analysis, the research is thus exploratory as the experimenters have not developed a falsifiable theory with associated predictions (Kennedy & Watt, 2018). One should identify a minimum effect size of interest that can be detected reliably, rather than based from previous studies. Identifying a minimum effect of interest and designing a study to assess whether the data is consistent with this effect is the best strategy for confirmatory research (Kennedy & Watt, 2018). Therefore, "before speculating about the theoretical reasons underlying the unreliability of evidence of ... parapsychological phenomena, we must exclude the possibility that it may be due to the neglect of statistical power" (Tressoldi, 2016, p. 14).

Frequentist or Bayesian? Both.

Frequentist methods can be designed to control (before data collection) the probability of error at the values of a parameter. Frequentists can be classified as prudent for their willingness to protect the error rate. However, frequentist methods do not allow for a posterior measure of belief. This posterior assessment is desirable and makes Bayesian analysis an attractive approach (Neath, 2010).

A major advantage to a Bayesian approach is the ability to assess certainty in an outcome (Neath, 2010). A Bayesian approach produces a distribution of plausible values for a parameter of interest, along with the associated probabilities (Tressoldi & Utts, 2015). However, a cost is that a Bayesian will not necessarily control the probability of error conditional on important possible parameter values. Further, Bayesian hypothesis tests are highly sensitive to the choice of a prior probability distribution and tend to be biased in favour of the null hypothesis, especially for small effects (Kennedy, 2015, 2016; Utts et al., 2010). If we think of frequentists as being prudent, Bayesians can be seen as impulsive for not protecting against these scenarios (Neath, 2010).

Operating characteristics are necessary for Bayesian hypothesis tests, as with classical hypothesis tests. The operating characteristics allow us to answer the question: If the true *ES* is a certain value, what is the probability that the planned analysis will give the correct inference? These evaluations quantify the expected rates of inferential errors and such evaluation establishes the statistical validity of a planned analysis (Kennedy, 2016). Evaluation of inferential errors and power quantifies the statistical validity of a planned hypothesis test. These evaluations determine the rates of correct and incorrect inferences if the true effect size is a certain value, and the corresponding rates if the null hypothesis is true.

For confirmatory research, a certain minimum effect size is often of interest and is the focus of the evaluation (Kennedy, 2015). Thus, the report must explicitly state all priors, including prior distributions represented in the alternative hypothesis (Tressoldi & Utts, 2015).

Statistical dependence

Statistical dependence is a problem within psychological and parapsychological experiments alike: human responses, be they conscious or physiological, cannot be assumed independent (Kennedy, 2016). This violates the assumption of independence, which is a common underlying principle for the dependent variable in statistical analysis. The impact of statistical dependence is difficult to predict but commonly makes *p*-values deceitfully small. Dependence problems are most likely to occur in studies with feedback to the participant on each trial, such as presentiment studies (Kennedy, 2016). Studies that use random events as the outcome variable are less affected, providing that the random events are truly random and thus are independent. Analysis that uses a binomial test with independent random events as the outcome variable are less likely to feature potential dependence problems (Kennedy, 2016).

2.6 Conclusion

Overall, there are numerous facets and factors to consider when planning a scientific experiment. Regulated medical research is often viewed as the gold standard to strive for, but many argue that achieving this level of rigour is unattainable in psychological research. As repeatedly demonstrated, findings in both psychology and parapsychology are potentially compromised by questionable research practices and a lack of transparency. The issues discussed in this chapter are not unique to parapsychology; they have significant implications for all experimental psychology research.

However, given the contentious nature of the phenomena investigated in parapsychology, concerns about security, fraud, and potential bias are routinely considered in the design of these experiments. This focus does not imply that parapsychological research is without flaws. The next chapter is a published manuscript of a meta-analysis intended to understand variability in study designs, specifically telepathy ganzfeld studies, which was intended to inform a telepathy study incorporating most of the recommendations stated in this chapter.

Chapter 3

Understanding the Factors at Play in the Sender-Receiver Dynamic During Telepathy Ganzfeld: A Meta-Analysis

3.1 Chapter Overview

The meta-analysis presented in this chapter began prior to the COVID-19 pandemic and was designed to inform the development of a telepathy ganzfeld study paradigm. At the time, little was known about the common features of telepathy ganzfeld studies, particularly regarding the sender-receiver relationship. Due to the lack of a defined theory, flexibility and 'ritualistic' properties of psi experiments (Rabeyron, 2020), this meta-analysis aimed to code existing telepathy studies to determine how prevalent certain features were and whether they influenced study outcomes. The goal was to use the findings to shape the planned telepathy ganzfeld design, incorporating as many anti-fraud measures and psi-conducive features as possible, as discussed in Chapter 2, section 2.3. By examining the features of previous telepathy studies, it highlights the variability in replication attempts and how small decisions can influence study outcome.

3.2 Published Manuscript

3.2.1 Abstract

Objective. To use meta-analysis to explore five previously uninvestigated factors related to the sender-receiver dynamic in the telepathy ganzfeld. The five factors of interest are: a) did the receiver see the sender's room prior to the session?; b) could the sender hear the receiver during the mentation period?; c) could the sender hear the receiver during the judging period?; d) was the sender explicitly told to be silent?; and e) did the experimenter

assist in the review section of the session? *Method*: Telepathy ganzfeld studies conducted post *Joint Communiqué*, with one session per day and the receivers rating the targets, were chosen. Two mixed-effects models were fit: 1) using the study hit rates as the binomial mean; and 2) using the study hit rates as a proportion. Both models have the five factors as binary moderators. *Results*: Both the binomial mean and proportion models suggest a significant effect of the moderators overall and two factors individually: 1) the sender being able to hear the receiver during the mentation period; and 2) a review period after the mentation period. Permutation tests for both models also show significant effects of the moderators and the two factors. *Conclusion*: The sender being able to hear the receiver's mentation appears to increase overall study success, while the review period decreases overall study success. This analysis is the first to examine the impact of these study design factors on the outcomes of ganzfeld telepathy experiments.

Keywords: meta-analysis, ganzfeld, psi, telepathy, extrasensory perception, anomalous cognition

Highlights

- Significant overall effect of the moderators on study success (hit rate)
- The sender being able to hear the mentation period was associated with a significant increase in study success
- A review period after the sending period was associated with a significant decrease in study success

3.2.2 Introduction

The ganzfeld is a procedure commonly used to test for anomalous cognition or extrasensory perception (ESP; Cardeña, 2018; Cardeña et al., 2015) and researchers have often reported replicable findings using this method (Baptista et al., 2015; Honorton et al., 1998; Storm & Tressoldi, 2020; Storm et al., 2010). The method uses an environment where the participant experiences a mild form of sensory deprivation. More specifically, the ganzfeld is defined as: homogeneous, unpatterned sensory stimulation: audio-visual ganzfeld may be accomplished by placing translucent hemispheres (for example, halved ping-pong balls) over each eye of the participant, with diffused light (frequently red in hue) projected onto them from an external source, together with the playing of unstructured sounds (such as "white" or "pink" noise) into the ears, and generally with the person in a state of bodily comfort; the consequent deprivation of patterned sensory input is said to be conducive to introspection of inwardly-

generated impressions, some of which may be extra-sensory in origin. [From the German for "entire field"]. (Parapsychological Association, 2015)

With a telepathy design, there are two participants, one acting as the sender and the other as the receiver. Telepathy can be formally defined as "Anomalous cognition (AC) to refer to ostensible acquisition of information in ways that are currently unexplained" with telepathy referring to the source presumably being another person's mind (Cardeña et al., 2015, p. 2).

During a telepathy ganzfeld session, the receiver is exposed to the ganzfeld environment. Their task is to become aware of the sender's thoughts while the sender views a randomly chosen target such as a video clip or static image in a different room. Usually receivers are asked to make a verbal report of any impressions or sensations they are experiencing and this mentation is audio recorded as well as being noted by an experimenter. Often the experimenter may review the mentation with the participant after the impression period, a part of the session typically referred to as the review period. After the impression (and review) period, the receiver views a random selection of decoy video/image clips, along with the target (the clip the sender was aiming to communicate). While the receiver and experimenter remain unaware of the identity of the actual target, the receiver ranks the similarity of their impressions with the presented targets. If the highest rated target is the same as the target that the sender was viewing the session is registered as a hit.

Parapsychologists have been reporting significant results in ganzfeld studies since the 1970s, however there has been little systematic investigation of which aspects of the experimental set-up are associated with elevated hit-rates. Ganzfeld design features such as target type have been analyed, with dynamic targets producing larger study effect sizes (Honorton et al., 1990), though this observation was not confirmed in Milton and Wiseman's (1999) analysis of "new generation" ganzfeld studies. Honorton (1977) reported that successful sessions have on average 37 minutes of ganzfeld exposure. Bem and colleagues (2001) found that more standard studies obtained higher hit-rates, although there is little consensus on the definition of the standard ganzfeld (Milton, 1999; Schmeidler & Edge, 1999).

Most attention has been paid to the role of the sender, with studies hoping to shed light on the sender's influence and whether it is instrumental (inherent to the communication process) or peripheral (pertaining to psychological or motivational factors). Honorton's (1995) meta-analysis of the ganzfeld literature reported that studies using senders perform better than

those without. However, later studies designed directly to compare sender and no sender conditions generally report no significant difference between conditions (e.g. Morris et al., 1995; Roe et al., 2003; Roe & Holt, 2005). Other potentially important aspects of the ganzfeld set-up, especially around the sender-receiver dynamic, remain unexplored. As Cardeña (2020) argues, although there is abundant evidence that the ganzfeld creates a psi-conducive environment, we need to be more systematic in investigating which elements of the ganzfeld procedure are important. Therefore, the focus of this meta-analysis is to explore how five previously uninvestigated ganzfeld telepathy study design features may influence study outcome.

Factors of Interest

Factor 1: Do Receivers See the Sender's Room?

Psi-conduciveness is often mentioned in the ganzfeld literature, with some stating that creating a warm and pleasant atmosphere creates a more psi-conducive session (Dalton, 1997; Milton, 1997a). However, more information is required about the detailed protocol of each study, especially concerning experimenter-participant interactions, which are quite extensive in the ganzfeld due to the one-to-one nature of the testing, and the duration of each session (M. D. Smith & Savva, 2004). Ganzfeld researchers often mention rapport building chats between experimenter and participants, however it is not known whether having the receiver see the sender's room before the session is a key aspect of the study. Being told there is a sender in a different room may be unnerving to a new participant, especially given the length and intensity of the study. Furthermore, perhaps any emotional or social connection between the sender and receiver will be stronger if the receiver has been introduced to the sender's environment and the sender, in turn, is aware that the receiver knows where they are. Thus, the rationale for assessing this factor is to understand whether scoring is higher when the receiver has seen the sender's room before the session commences.

Factor 2: Do Senders Hear the Receiver's Mentation Live?

It takes some effort to set up a one-way audio connection from the ganzfeld receiver to the sender. Researchers often make this effort thereby enabling the sender to hear the receiver describe out loud their feelings, impressions and sensations during the sending period. The stated justification for this design feature is to allow the sender to mentally reinforce the correct images and impressions to the target (Dalton, 1997) and to add an air of

excitement and active involvement for the participants (Parker et al., 1997). If the sender's influence is peripheral via motivational factors (as suggested by Honorton, 1995), then the sender being able to hear the receiver during the sending period may be the most important aspect of a telepathy design because by hearing the receiver's mentation the sender should, theoretically, be able to reinforce receivers at times when they seem to be describing the target. Further, not every ganzfeld study features this audio channel (or perhaps does not clearly report this aspect of procedure), so it is also of interest to establish how common this practice is.

Factor 3: Do Senders Hear the Receiver During the Judging Period?

If the sender can hear the receiver produce their mentation during the sending period and reinforce impressions linked to the target (Factor 2), the same logic extends to the judging period - the time when the receiver decides which target clip is most like their experiences. As the receiver (and/or the experimenter) are reviewing the mentation report and making decisions about the ratings, the sender being able to hear this should, in theory, be able to mentally reinforce the target. Nevertheless, to date there has been no analysis of the impact of the sender hearing the judging period.

Factor 4: Sender Told to be Silent

This factor is primarily assessed from a target security concern: even if senders are physically distanced from the other experimenters and receiver and the target is shielded, explicitly telling the sender to be silent provides an extra layer of security. Although a minor aspect of the ganzfeld procedure, if this factor is significantly related to study outcome then it suggests that previous studies may have been susceptible to sensory leakage. Instructions for the sender to "silently communicate" the target have been used in study protocols throughout the ganzfeld literature (e.g., Berger & Honorton, 1986; Honorton et al., 1990). Thus, assessing the prevalence of this factor may help us to evaluate whether subtle sensory leakage (such as vibrations) may potentially influence study outcome, even if acoustic shielding in these studies is assumed to be adequate (see Wiseman et al., 1994).

Factor 5: Mentation Review

It is common for ganzfeld studies to have review periods (after the sending and before the judging period) in which the experimenter reviews the mentation notes and allows the receiver to elaborate or clarify their mentation (Kanthamani & Broughton, 1994; Roe et al., 2004; Watt et al., 2020). The review period may assist participants in processing their

experiences, remembering their mentation and in making connections between their mentation and the targets that they may have otherwise not noticed (Wooffitt, 2003). However, there has not yet been a systematic review of the importance, or otherwise, of the mentation review.

Objective

This meta-analysis is exploratory as there has been no previous systematic review of the above aspects of the ganzfeld study procedure. The research questions originate from a pragmatic motivation: to provide evidence to guide the design of future ganzfeld telepathy studies. Hence, there are no expectations from the analysis. Nonetheless, the null hypothesis is that there will be no effect of the moderators (five factors) on study success (hits significantly greater than chance). Ganzfeld telepathy studies published between January 1988 and September 2021 are included, to assess studies conducted with the potential benefit of the methodological guidelines from the *Joint Communiqué* (Hyman & Honorton, 1986).

The independent variables are the five factors (detailed above) rated by two raters (details in Appendix A). The dependent variable for both meta-analytic models is the study hit rate (percentage of hits). The first author created two models, the first treating the study hit rate as a mean, following the approximated binomial distribution. The intention was to use the *z*-scores for each study, which are approximated from the binomial distribution, but the standard deviations could not be computed. The second, supplementary, model uses the study hit rate as proportion. Homogeneity analysis was automatically calculated by the model function, which calculated the I^2 value. The I^2 statistic describes the variation across studies due to heterogeneity, rather than chance and is a simple description of the inconsistency of studies' results (Higgins et al., 2003; Higgins & Thompson, 2002). Analyses were conducted with RStudio Workbench Version 1.4.1717-3 © 2009-2021 RStudio, PBC, and models built with the *metafor* package (Viechtbauer, 2010). A protocol was not pre-registered for this project due to its exploratory nature. The data set used for analysis, analysis code, model formulae and project information are all publicly and freely available at https://github.com/yeloopa/telepathyMA.git

3.2.3 Method

Inclusion and Exclusion Criteria

We include all ganzfeld telepathy studies using visual targets and human participants from January 1988 to September 2021. All studies had a measure of hit rate (%) as well as session binomial *z*-score as their outcomes, and a four-option design with one target and 3 decoys (thus resulting in a 25% mean chance expectation; MCE). Study hit rate is a percentage calculated as the overall number of hits obtained across the study sessions. The associated *z*-scores are the related binomial distribution *z*-ratio for situations of the general "k out of n" type.

Information Sources

We first extracted all telepathy studies from Tressoldi's ganzfeld database, accessible at the *Society for Psychical Research's* open-access data website *Psi Open Data* (Tressoldi, 2019), which contains all studies conducted since 1974 to 2018, used for the two recent meta-analyses of the ganzfeld literature (Storm et al., 2010b; Storm & Tressoldi, 2020).

To check for studies produced since 2018, a literature search on Google Scholar was conducted, using the search terms "ganzfeld," "telepathy," and "study," using the Boolean connector "&" in the title and abstract fields for the years 2018 to 2021. Inspection of reference lists of included papers was also used as a part of the search strategy to ensure all relevant studies were included. The literature search resulted in one addition, a telepathy study published in 2020 (Cardeña & Marcusson-Clavertz, 2020). ALP contacted the author of the 2020 paper to establish if the multiple sessions had been performed on a single day by the participants – which they had not.

Study Selection

The study selection procedure is outlined in Appendix A. Studies were excluded based on these criteria:

- Duplicated: for example, if a published paper was produced from a conference proceeding, the conference proceeding was removed from the database. Published papers have more detailed and full analyses and usually all planned sessions are complete.

- Studies using external judges: the factors Hear and Hear judging asses if the sender hearing the receiver during the two periods (outlined above) influences study outcome. Thus, using external judges would not help this assessment, especially for the Hear judging factor (see Hyman, 1995).

- Multiple sessions a day: if the studies had a repeated participants design which ran multiple sessions on the same day, they were removed. Participants contributing to multiple sessions violates the independence assumption of most statistical hypothesis tests. Likewise, there is literature noting a decline effect due to fatigue in experimenters (Broughton & Alexander, 1997; Parker et al., 1998; Wezelman & Bierman, 1997)
- Multiple trials per session: like the point above, only one trial per session in the current study design.
- Stimuli material: Studies which used non-visual targets were removed. Visual targets are considered standard (specifically dynamic targets; Bem et al., 2001).
- Multiple or no senders: Some studies included designs that involved 0, 1, or 2 senders. In these reports, it was unclear if the analysis combined all of these trials into one analysis, or the different sender options had low sample sizes (< 10).
- Low sample size: studies with samples of 10 or less were removed because of potential bias in results stemming from sampling error.

The first author conducted an initial analysis presented at the *Society for Psychical Research's* 2021 Conference (Pooley, 2021) using the data set composed of the studies rated by her using the study hit rate as the outcome measure. She stated then that this was not the final analysis, with corrections and changes still to be made, such as resolving the rater discrepancies and analysis with *z*-score as an outcome. After the conference and study selection was assumed to be final, numerous reports in the *Journal of Scientific Exploration* reported the serious fraudulent actions and widespread plagiarism conducted by Alejandro Parra (Braude, 2021; Cardeña, 2021; Nahm, 2021). Because of the seriousness of the accusations and evidence collected, we deemed it best to remove all the Parra studies from

the dataset (5 data points). Likewise, there was duplicate reporting of some results in the Gothenburg study series (Parker et al., 1997; Parker & Westerlund, 1998). Given the removals and corrections to the database, the final data set is 41 studies.

Data Extraction and Coding

As the data are primarily sourced from a freely available ganzfeld database (Tressoldi, 2019), the number of variables of interest for each experiment were reduced to:

- Study author(s) and year (and series number if multiple series per study)
- Study hit rate (%)
- Study *z*-score
- Number of participants
- Number of trials
- Studies were then organized according to the presence of the five factors of interest:
- Did the participant (receiver) see the sender's room before the session? (See)
- Could the sender hear the receiver produce their mentation? (Hear)
- Could the sender hear the receiver during the judging period? (Hear judging)
- Was the sender explicitly told to be silent? (Silent)
- Did the experimenter review the receiver's mentation notes with the receiver, after the sending period? (Review)

Each factor is rated on a binary scale (0 for no presence, 1 for presence of factor). See Appendix A for the instructions given to the raters. For the five factors, the first author first assessed each paper and provided ratings based on the instructions. A second rater then did the same, following the same instructions. However, because of health issues with the second rater, it was not possible to arrange a meeting to resolve the discrepancies in ratings and those ratings were disregarded, so CW was recruited as a third rater. Discrepancies in ratings between Raters 1 (ALP) and 3 (CW) were resolved in a meeting (see Table 1 for the inter-rater reliability scores). Due to the ambiguity of what constitutes a "review period," ALP and CW agreed in the rating meeting that if the study report explicitly stated a review occurred, then it was rated a 1. If a review period was not explicitly stated in the paper, it was up to the rater to decide if a review stage could be inferred: hence the rating instructions included the

opportunity for the receiver to add, alter and/or discuss their mentation with the experimenter (see Appendix A).

Summary Measures

Primary Analysis: Binomial Test with Mean Number of Hits

For the primary analysis we intended to use the *z*-score as the outcome measure. However, we noted that a recent pre-publication of a ganzfeld meta-analysis has been criticized during open peer review for the effect size calculation for the study *z*-scores as not being scientifically sound (see Tressoldi & Storm, 2021b). Thus, for the current analysis, ALP performed the Binomial test using the mean number of hits rather than the total number of hits, using a random-effects model with study hit rate treated as a mean. The five factors (all binary) of interest were added as moderators in the model, thus resulting in a mixed-effects model. The final model is as follows:

$$\theta_i = \beta_0 + \beta_1 * See + \beta_2 * Hear + \beta_3 * Hearjudge + \beta_4 * Silent + \beta_5 * Review + \mu_i$$

where,

$$\mu_i \sim N(0, \tau^2)$$

First, the number of trials, study hit rate and the associated standard deviation of the binomial distribution for treating the hit rates as a mean were calculated and entered into the *escalc* function, which resulted in the observed effect sizes and sampling variances in order to fit the meta-analytic model. We estimated the heterogeneity of the effect sizes by fitting the model with a restricted maximum-likelihood (REML), which is better when working with smaller samples (Viechtbauer, 2005), and using the Knapp-Hartung adjustment (Knapp & Hartung, 2003).

Secondary Analysis: Proportion of Hits

For this model, we used the proportion of hits as the outcome measure. As the meta-analysis used aggregate scores that provide data about individual groups in respect to a dichotomous dependent variable (hit or miss), the number of events and number of trials are required to calculate the appropriate effect size. Due to the similarity between the primary and secondary models, we report the primary analysis in full in the Results section with the model prior to the removal of outliers. The secondary model results can be found in full in Appendix A.

Methods of Synthesis

Because of the exploratory nature of the study and lack of previous relevant research, a meta-regression for both models was created. Meta-regression not only provides a summary of the selected studies but also evaluates how the five potential moderators may influence study outcome. Study results are not weighted.

Publication Bias and Selective Reporting

Funnel plots for each analysis were created (a funnel plot is a useful visual aid to assess publication (and other) potential bias in the database). Because of the inclusion of moderators in the models, we could not perform the trim-and-fill method to assess publication bias. However, multiple reviews of the ganzfeld literature have reported no suggestion of publication bias problem (Baptista & Derakhshani, 2014; Cardeña, 2018; Storm et al., 2010b).

3.2.4 Results

Inter-Rater Reliability

For the sake of transparency, Table 1 presents the results of the initial ratings between Raters 1 and 3. These disagreements were resolved before the final analyses reported below (also see Appendix A).

Table 1

Unweighted Kappa Scores and Observed Agreement Between Rater 1 and Rater 3

Factor	Kappa	Observed agreement
See	76%	83%
Hear	50%	78%
Hear judging	21%	71%
Silent	33.5%	63%
Review	8%	71%

Descriptive Statistics

Appendix A details the studies included in the models and the respective measures. A total of 41 studies (or series reported as part of a wider study) were conducted by 17 different lead authors who reported their results in a total of 23 articles. All 41 studies used a four-choice design therefore the mean chance expectation is 25%. A total of 1,496 participants contributed to 1,624 ganzfeld telepathy sessions. The average hit rate across the studies is 32% (*SD* = 10.40%) with a skewness of 0.72. An exact binomial test is reported to assess if the mean hit rate in the data set is significantly greater than chance. There are 520 hits in 1,624 trials, resulting in a significant difference from chance at the 5% significance level: Binomial Exact $p < .001$, one-tailed[3]. The mean *z*-score is 0.91 (*SD* = 1.37), the sum of *z*-scores is 37.20 and Stouffer's *Z* is 5.81[4].

Results of Binomial Mean Model (Model 1)

The model was first fit with all 41 studies. The model output is reported in Table 2. The test for moderators is non-significant at the 5% level $F(5, 35) = 1.89$, $p = .12$. However, the coefficient Review is significant suggesting that the presence of a review session in a ganzfeld telepathy study decreases the average study success by 12%, when all the other study features are set to 0. See Table 2)

[3] A 'greater' alternative was used given the ganzfeld literature showing that the average hit rate in ganzfeld studies is around 32%.

[4] The *z*-score descriptives are provided for the sake of comparability to other meta-analyses of the ganzfeld literature. As discussed earlier, the *z*-scores were not used in the models reported.

Table 2

Model 1: Hit rate as Binomial Mean Summary Output Prior to Influential Studies Removed

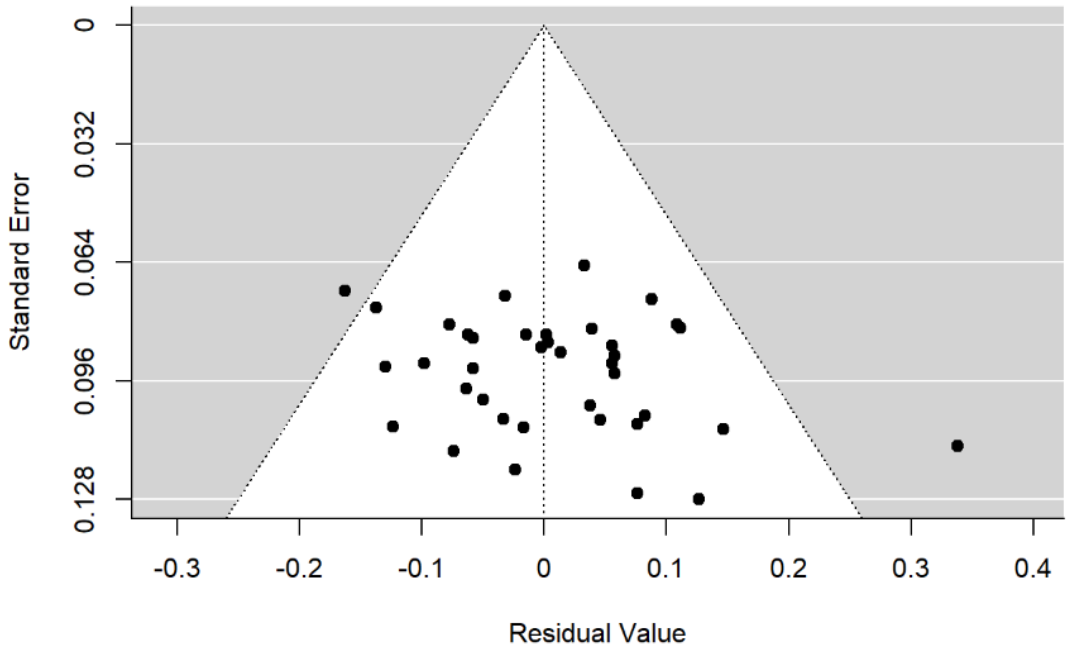
	Estimate	Standard error	*t*-value	*p*-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.38	.07	5.71	<.0001***	0.25	0.52
See	-.01	.05	-0.32	.75	-0.11	0.08
Hear	.07	.04	1.66	.11	-0.01	0.15
Hear judging	-.02	.04	-0.35	.73	-0.07	0.09
Silent	.01	.04	0.20	.84	-0.07	0.09
Review	-.12	.06	-2.16	.04*	-0.24	-0.01

Note. * indicates significance at the 5% level, *** indicates significance at the <.0001

The unaccounted variability in the model is moderate (I^2 = 43%) and the QE test for residual heterogeneity is significant $QE(35) = 62.38$, $p = .003$. As shown in Figure 1, the funnel plot suggests that there is publication bias in the dataset. A mixed-effects meta-regression model Egger's regression test for funnel plot asymmetry was performed and was significant $t(34) = 2.15$, $p = .04$. Further assessment of the model revealed there were influential studies in the dataset. First, Honorton 302 (Honorton et al., 1990) was removed due to having the highest standardized residual value of 3.41, where standardized residuals between -2 and 2 are commonly used as acceptable limits. The model with Honorton 302 removed still flags an influential case, Goulding et al. (2004) with a standardized residual value of -2.7. With the removal of Honorton study 302 and Goulding et al. (2004), the model checks flag another influential study exceeding the limit: Broughton and Alexander (1997) FT2 has a standardized residual value of -2.7. The model was run once again and checked for influential values and returned no more influential cases, thus we detail the final model below.

Figure 1

Funnel Plot for Model 1 Prior to Removal of Influential Studies



Binomial Mean Final Model (Model 1.1)

With the three influential cases removed, the test for moderators is significant at the 5% level $F(5,32) = 3.78$, $p = .01$. As in Model 1, Review is significant and Hear now is also significant, as shown in Table 3. On average, study success rate is increased by 7% when the sender can hear the mentation period, when all other factors are set to 0. However, the addition of a review period in a ganzfeld telepathy study decreases the average study success by 10% when all other factors are set to 0. A permutation test (5000 iterations) was performed and confirmed the findings, with the test for moderators significant $F(5,32) = 3.78$, $p = .016$. The factors Hear (Factor 2) and Review (Factor 5) were significant, as shown in Table 4. The forest plot for the final model (Model 1.1) is shown in Appendix A.

Table 3

Model 1.1: Hit Rate as Binomial Mean Summary Output

	Estimate	Standard error	*t*-value	*p*-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.36	.05	7.22	<.0001***	0.26	0.46
See	.00	.03	0.10	.92	-0.07	0.07
Hear	.07	.03	2.26	.031*	0.01	0.13
Hear judging	-.04	.03	-1.10	.28	-0.11	0.03
Silent	.02	.03	0.62	.54	-0.04	0.08
Review	-.10	.04	-2.40	.022*	-0.19	-0.02

Note. * indicates significance at the 5% level, *** indicates significance at the 0.1% level.

The estimated amount of residual heterogeneity for the model is very low ($\tau^2 < .001$, $SE = 0.001$), unaccounted variability is also low ($I^2 = 2.8\%$) and total amount of heterogeneity accounted for by the model is very high ($R^2 = 92\%$). Model funnel and forest plots are shown in Figures 3 and 4, respectively.

Because of the inclusion of moderators in the model, the trim-and-fill method could not be performed[5]. However, as shown in Figure 3, there is little evidence to suggest publication bias in the data set. Furthermore, a mixed-effects meta-regression model Egger's regression test for funnel plot for asymmetry was performed and was non-significant $t(31) = 0.17$, $p = .87$.

Binomial Mean Model with Review Factor Removed (Model 1.2)

Because of the high incidence rate of the Review factor (only 3 studies were rated to have *no* identifiable review period), the Binomial final mean model (Model 1.1) was performed with the Review factor removed to assess if the model results changed with this factor removed. The model formula is the same as the primary analysis, just with this factor removed. Given the similarity between the Binomial mean model (Model 1.1) and Proportion of hits model

[5] The trim-and-fill method was applied to the binomial mean model without the moderators, using the original dataset and the dataset that had influential studies removed. Both trim-and-fill analyses estimated no missing studies.

(Model 2), only the Binomial mean model was performed. Full results are reported in Appendix A.

Table 4

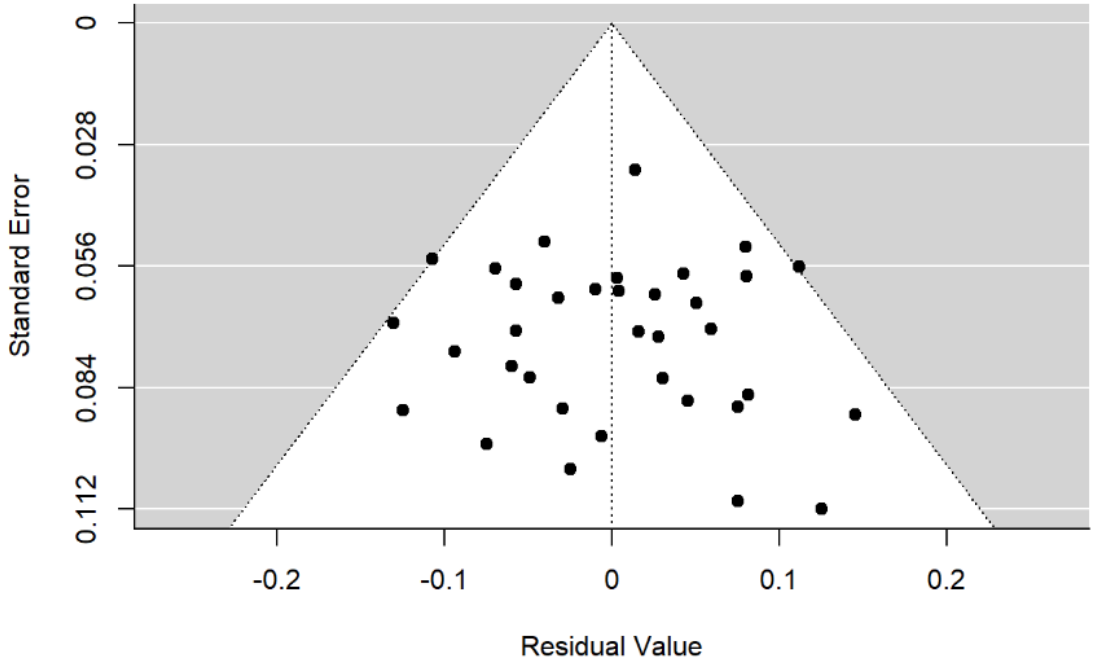
Permutation Test Results (5000 Iterations) for Model 1.1

	Estimate	Standard error	*t*-value	*p*-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.36	.05	7.22	.04*	0.26	0.46
See	.00	.03	0.10	.92	-0.07	0.07
Hear	.07	.03	2.26	.04*	0.01	0.13
Hear judging	-.04	.03	-1.10	.29	-0.11	0.03
Silent	.02	.03	0.62	.55	-0.04	0.08
Review	-.10	.04	-2.40	.03*	-0.19	-0.02

Note. * indicates significance at the 5% level.

Figure 2

Funnel Plot for Model 1.1



3.2.5 Discussion

In this study, we used a meta-analysis to explore how five potentially important aspects of ganzfeld telepathy study procedure might be associated with study outcome and found two factors that had significant impacts on study outcome. First, studies that allowed the sender to hear the receiver's mentation (Factor 2) were associated with an increase in study hit rate by approximately 7%. This suggests that the sender hearing the receiver may motivate them to actively reinforce the target during the sending period and keep the receiver "on track". An alternative interpretation arises from work by Fox (2004), who found senders are susceptible to wandering and boredom, suggesting that perhaps an audio link is sufficient reinforcement to keep the sender engaged and motivated throughout the session.

In contrast, the inclusion of a mentation review period (Factor 5) was associated with a significant decrease in study hit rate by 10%. This supports Dalton's observations in her doctoral thesis that the experimenter-receiver interaction during the review period decreases study success (Dalton, 1997). However, this finding is based upon only three studies (after influential cases were removed) that the raters judged did not have an identifiable review period. Nonetheless, perhaps the review period with the experimenter and receiver discussing correspondences between the mentation and clips does not provide clarity but rather introduces ambiguity and could allow the experimenter to direct the receiver away from the target. Alternatively, perhaps the review period introduces confusion about who (experimenter or receiver) is making the final judgement. In support of this latter interpretation, Model 1.2 with the Review factor (Factor 5) removed resulted in Hear (Factor 2) becoming more significant than in Models 1.1 and 2 and Hear judging (Factor 3) becoming significant (as shown in Appendix A).

Our analysis revealed two factors to be clearly unrelated to study outcome. Factor 1 (See) was not significant in any model nor near significance at any point, suggesting that knowledge of the sender's location may not have a major influence over telepathy study success. Likewise, the non-significance of Factor 4 (Silent) may suggest that sensory leakage is less of a concern than has been suggested for some of the early ganzfeld telepathy literature (Wiseman et al., 1994), but this factor is an indirect measure of sensory leakage.

One caveat to our finding that the mentation review is associated with lower hit-rates, was that the rating instructions did not distinguish between a review period which was stipulated in the procedure in order for receivers to clarify or elaborate upon their mentation,

but a simple opportunity for the receiver to add, comment upon, or discuss their mentation period. This is in contrast to a scenario where the receiver moves immediately from the sending/mentation period to the judging period. Hence, there is scope for future research to look more closely at the researcher-receiver relationship and to investigate how interactions during the session may influence study outcome. Likewise, the inter-rater reliability (kappa) scores are noticeably varied: Factor 5 (Review) was significant in all models, but this factor had the lowest kappa values. This illustrates a limitation with kappa values: kappa is not reliable for rare observations and low values of kappa may not necessarily reflect low agreement overall (Viera & Garrett, 2005), and given that there were four studies rated as not having a review period, the kappa value is unsurprisingly low for this factor. During the meeting to resolve disagreements, both raters discussed their ratings. Some discrepancies were merely mistakes whereas others were different interpretations. For example, for the Hear judging factor, it was decided during the rating meeting that if (after the receiver had logged their ratings) the sender was summoned via the one-way audio link then this *implied* that the sender could hear the whole judging procedure (given they could already hear the mentation period). Hence, the ambiguity of the phrasing used by the authors in their study methods allows for different interpretations of the study designs and we recommend future ganzfeld researchers take care to give a comprehensive description of possibly important aspects of study procedure. Similarly, with the Review factor, the rater clarified during their meeting that if there was an *opportunity* for the receiver to add, comment, or discuss their mentation notes with the researcher then this was rated as a review period, as were the more clearly defined review periods where the receiver *had* to explicitly elaborate their mentation and experiences during the sending period. This a limitation with the rating instructions due to the broad criteria for the Review factor.

One weakness of the current analysis is that the studies are not rated or weighted in terms of their study quality. For the current report, only studies produced years (1988 onwards) after the *Joint Communiqué* (Hyman & Honorton, 1986) were included, in an attempt to exclude earlier studies that did not have the benefit of Hyman and Honorton's methodological recommendations. Nonetheless, studies with better security protocols and clearer method sections could have been given more weight in our meta-analysis.

Even so, there are still some valuable findings from our analysis. First, this meta-analytic review shows that the results found in both models (Model 1.1 and Model 2) were confirmed in the permutation tests of the model coefficients. The omnibus test for the

moderators shows that there was a significant effect of the factors, with the Hear and Review factors significantly affecting study outcome. Second, even when using different outcome and effect size measures than previous meta-analyses (Storm et al., 2010b; Storm & Tressoldi, 2020), the results previously reported in the literature still stand: the noise-reducing ganzfeld protocol significantly produces hit rates greater than MCE. This is not surprising as the studies we used were primarily extracted from the same database; however, there is no need to account for selected participants in the current analysis as heterogeneity is low in both models, unlike what Storm and colleagues (2010b) found. Third, although our study is limited in its generalizability, it has provided a new angle to look at the telepathy ganzfeld literature and can perhaps aid in providing an evidence-based procedure for future ganzfeld telepathy studies. The vast majority of reported ganzfeld telepathy studies have a review period, even though it appears from our analysis to be detrimental to study outcome. Likewise, the difficulty we had in coding some study reports should encourage future researchers to provide more detail when reporting their study designs. Given current day open science infrastructure and internet-based dissemination possibilities available, there are plenty of opportunities for full disclosure of study designs. This meta-analysis also adds to recent publications looking at study design factors in the ganzfeld, such as Schmidt and Prein's (2019) study investigating different auditory homogenizations and Kübel et al.'s (2021) assessment of red vs green light visual stimulation. This suggests that there is interest in ganzfeld design factors and prospective ganzfeld researchers should try to make their methodological decisions based on the available evidence.

3.3 Conclusion

The meta-regression revealed that allowing the sender to hear the receiver had a significant positive effect on session outcomes, while having the experimenter discuss the mentation notes with the receiver significantly decreased session outcomes. These findings were not known before this analysis and led to changes in the mentation review period for KPU Study 1074. Although this analysis was conducted prior to a shift in the thesis direction due to pandemic-related experimental delays, it underscores how little is understood about the core features of ganzfeld paradigms and highlights that not all ganzfeld studies share the same foundational elements. This chapter demonstrates the flexibility and researcher degrees of freedom involved in conducting an experiment, using the telepathy studies as a microcosm. The next chapter addresses the flexibility and researcher degrees of freedom

involved in conducting meta-analyses, arguing that meta-analyses are far from being a "controversy killer." It uses the example of psi ganzfeld meta-analyses to illustrate the inherent issues with retrospective meta-analyses.

Chapter 4

Methodological umbrella review of psi ganzfeld meta-analyses

To do a meta-analysis is easy; to do a meta-analysis *well* is not

(Sharpe & Poets, 2020, p.382)

4.1 Chapter overview

After conducting my own meta-analysis, presented in Chapter 3, I gained first-hand experience into the ambiguity and numerous decision points over which a researcher has control during the meta-analytic process. This experience highlighted concerns about research degrees of freedom, prompting a deeper exploration of the issues discussed in Chapter 1 regarding QRPs and the limitations of conventional meta-analyses, as highlighted in Chapter 2. Given the long and contentious history of psi ganzfeld meta-analyses, I used this literature as an example to examine how methodological choices—such as the use of post-hoc data and lack of transparency regarding alternative analytic pathways—can influence outcomes. This current chapter presents a methodological umbrella review designed to investigate the variability in how researchers conduct meta-analyses of psi ganzfeld studies, with a particular focus on the inclusion criteria they report.

4.2 The problem with meta-analyses

As briefly defined in Chapter 2, meta-analysis is a commonly used method in psychology to quantitatively summarise numerous research publications focusing on a similar topic, methodology or research question. A meta-analysis can be useful as it allows inferences form a larger pool of samples and encourages systematic research (Nelson et al., 2018). As defined in Chapter1, meta-analyses are also prone to questionable research

practices like primary research (Voracek et al., 2019; Wagenmakers et al., 2011; Watt & Kennedy, 2017) and may even exacerbate problems in primary research, such as *p*-hacking, reporting errors and fraud, whilst also overestimating effect sizes (Bartoš et al., 2023; Nelson et al., 2018; Sladekova et al., 2023). Further, both meta-analyses and primary research often neglect to address the flexibility in data-analytic decisions, including the selection, coding, organisation and analysis of data (Steegen et al., 2016; Taylor & Munafò, 2016; Voracek et al., 2019), commonly referred to as researcher degrees of freedom (Wicherts et al., 2016). Research articles often give inaccurate information about how researchers developed their hypotheses, analysed their data and drew conclusions (Schwab & Starbuck, 2017). As authors of meta-analyses frequently do not describe their process of hypothesis generation, leading to a potential bias when conducting electronic literature searches for a meta-analysis. Bosco et al. (2016) warns that if studies supporting these hypothesised relationships are more likely to be published, then these relationships are reflected in the titles and abstracts of studies, increasing the risk of upwardly biasing estimates. This raises concerns about the potential impact of under-powered and poorly designed primary studies, as well as the subjectivity of procedures and interpretation when constructing a meta-analysis (Kennedy, 2013; Murray, 2011; Nelson et al., 2018). Although reporting guidelines exist for meta-analyses and systematic reviews (e.g., PRISMA, MARS), they still do not shed light on decisions surrounding inclusion criteria and which modelling to use, nor the alternative ways the meta-analysis *could* have been constructed.

Likewise, quality assessment tools for meta-analyses are not commonly employed in psychology. While risk of bias assessments are more routinely used, quality assessment tools remain underutilised (Oliveras et al., 2017). For instance, a review of industrial and organisational psychology found that 98.2% of 120 systematic reviews and meta-analyses did not assess study quality, despite the introduction of reporting guidelines (Schalken & Rietbergen, 2017). Similarly, a review of health and clinical psychology studies found that only 19% related the outcomes of study quality assessments to the results of the consequential review (Oliveras et al., 2017). Given the overall low transparency in psychology reviews (Polanin et al., 2020; Sharpe & Poets, 2020), it is difficult to gauge if and what methodological quality tools are being used. Although this issue goes beyond the scope of the current project, which focuses on inclusion criteria, it is important to emphasise that the use (or lack) of quality assessment and risk of bias tools is another facet of researcher degrees of freedom.

This chapter briefly summarises the literature on researcher degrees of freedom and the acknowledged limitations of post-hoc analyses in psychology research, as defined in Chapter 1. Researcher degrees of freedom as measured by study inclusion criteria and how these decisions have shaped meta-analyses looking using psi ganzfeld data. As alluded to in Chapter 1, the ganzfeld debate exemplifies a subject marked by conflicting meta-analyses that fail to resolve issues inherent in primary studies, perpetuating an unproductive debate (Voracek et al., 2019). Further, the psi ganzfeld is notable not only for the controversy surrounding the topic but also as it stands out as one of the few areas characterised by extensive discourse on the use of meta-analysis (Milton, 1999; Schmeidler & Edge, 1999). While meta-analysis is generally perceived as an authoritative tool for objectively assessing the strength of evidence (Taylor & Munafò, 2016), those produced about the psi ganzfeld debate often provoke debates and discussion, especially regarding the methodological and statistical practices employed. Consequently, the psi ganzfeld meta-analysis serves as an exemplar of the issue of researcher degrees of freedom in methodological and statistical choices, which is actively (and publicly) scrutinised by others. This challenges the notion that meta-analyses have the conclusive "final word" in academic debates (Taylor & Munafò, 2016). Thus, using a methodological umbrella review, this chapter aims to clarify the construction of meta-analyses in the psi ganzfeld literature and investigate how these constructions shape the interpretations of their findings.

4.2.1 Researcher degrees of freedom

Since the replication crisis in psychology (c. 2011), there has been increasing focus on the fragility of findings and the flexibility when selecting, analysing and reporting data, often known as researcher degrees of freedom (Gelman & Loken, 2016; Simmons et al., 2011; Wagenmakers et al., 2011; Wicherts et al., 2016). Wicherts et al. (2016) suggest that there are at least thirty-four degrees of freedom researchers face when generating a hypothesis, designing, collecting, analysing and reporting psychological research and most of these decisions are interrelated. For example, in the early stage of a research project, such as generating a hypothesis, if a researcher conducts exploratory research without explicitly hypotheses, the decision may impact latter stages such as reporting their results, especially if the researcher presents this exploratory analysis as confirmatory (also known as ’hypothesising after results known’ or HARKing; Kennedy, 2013; Wicherts et al., 2016).

Early decisions in the meta-analysis process, such as which data to analyse and inclusion criteria have the most influence on meta-analytic outcomes, rather than the statistical models (Goodyear-Smith et al., 2012; Voracek et al., 2019). However, Wicherts et al. (2016) also note that the decisions psychological researchers face are often arbitrary and terms such as "questionable research practices" may be too harsh for the average psychologist. Nonetheless, Simmons et al. (2011) argue that this ambiguous and exploratory behaviour is commonplace and driven by two factors.

First, the uncertainty about how best to design and analyse a study and second, the researcher's desire (knowingly or not) to produce a statistically significant result. Hence, the problem is not that the average researcher tries to deliberately obtain favourable results, but rather they have a strong potential and opportunity to create bias (Wicherts et al., 2016).

One advocated practice to reduce bias is preregistration. This was promoted soon after the replication crisis as an easily implementable practice to create greater transparency in these flexible data-analytic decision points (John et al., 2012; Steegen et al., 2016; Wagenmakers et al., 2011). Indeed, preregistration is increasingly popular for both primary studies and meta-analyses (Lindsay & Nosek, 2018; Simmons et al., 2021) but is lagging for systematic reviews (Ioannidis, 2016). Whilst preregistration may improve clarity between exploratory and confirmatory research, it does not inherently lead to better theories, methods or analyses (Ioannidis, 2016; Lindsay & Nosek, 2018; Szollosi et al., 2020). More recently, preregistration is being discouraged by some due to the seemingly arbitrariness of preregistering studies for the sake of it and potentially causing more harm than good (Szollosi et al., 2020). Preregistration does not increase transparency about the multitude of decisions researchers make when designing a study, or even when preregistering one as "manoeuvrability remains if preregistrations are not sufficiently specific, precise or exhaustive" (Wicherts et al., 2016, p. 2). Thus, reported analyses may not be representative of possible valid specifications and may be biased by the hidden use of degrees of freedom, and this is true even if studies are preregistered (Del Giudice & Gangestad, 2021).

4.2.2 Post-hoc analysis

As defined in Chapter2, retrospective meta-analyses are a form of post-hoc analysis as the researchers are already aware of the findings of primary level studies before including (or

excluding) them in their meta-analysis (Kennedy, 2013). There is a greater risk of effects being biased upwards when conducting electronic database searches for a meta-analysis due to HARKing, which promotes larger and significant studies over smaller and non-significant findings (Bosco et al., 2016). The use of post-hoc analysis is often criticised as a form of 'dredging' the data in the hope of finding something worthy of publication (Srinivas et al., 2015). Knowing the results of primary level studies when conducting a meta-analysis, the researcher may (un)knowingly bias their selection criteria, model specification, quality ratings and so on, to create a preferred outcome (Kennedy, 2013). Even with greater transparency and strict protocols when performing systematic reviews, seemingly objective tasks become increasingly flexible and subjective (Goodyear-Smith et al., 2012; Ioannidis, 2016). Goodyear-Smith et al. (2012) highlight research will be influenced by researchers with strong prior belief and shape results depending on their expectations. Thus, strong prior beliefs in combination with post-hoc data usage will further exacerbate the subjectiveness of meta-analytic decisions. As decisions about which studies to include are the most influential, whilst also being the most ambiguous when designing a meta-analysis, umbrella reviews and other qualitative systematic reviews should examine the content specific inclusion criteria and data inclusion decisions used by the researchers.

4.2.3 The psi ganzfeld debate

As detailed in Chapter 1, meta-analysis is commonplace in parapsychology and especially those assessing the psi hypothesis via the psi ganzfeld paradigm. However, even with the same database and research question assessing if psi exists, researchers have come to diverging findings, as demonstrated by Hyman (1985)'s null findings versus Honorton (1985)'s positive findings. Since then, numerous meta-analyses and summaries of the literature have proclaimed there remains replicable evidence of the psi ganzfeld, even after tightening of anti-fraud methods (Bem et al., 2001; Bem & Honorton, 1994; Cardeña, 2018; Storm et al., 2010b; Storm & Tressoldi, 2020). Other researchers, however, have found a lack of replication (Hyman, 2010; Milton & Wiseman, 1999; Rouder et al., 2013). Regardless of the outcome of an individual meta-analysis, a debate ensues following its publication, with, proponents asserting that positive findings demonstrate evidence of psi, whilst sceptics scrutinise the work for methodological and statistical weaknesses, and vice versa for null findings (Wiseman, 2010).

As previously stated, the psi ganzfeld is characterised by an active discourse among researchers, who frequently engage in public discussions and critiques of recently published meta-analyses, regardless of their findings. This discourse is exemplified in the edited volume by Schmeidler and Edge (1999), which comprises of email discussions on various aspects of psi ganzfeld research, including the purpose of meta-analysis, relevant inclusion criteria, data analysis methods, and even the definition of the ganzfeld method.

The extensive discourse among parapsychologists was sparked by a null finding presented by Julie Milton in 1997 at the Parapsychological Association annual conference and later published (Milton & Wiseman, 1999). Their findings challenged the claim that the psi ganzfeld is a replicable paradigm across multiple experimenters under stringent methodological conditions. According to Milton (1999), after presenting the null results at the conference multiple researchers in the field recommended different methods for calculating or accumulating individual study outcomes, as well as various criteria for identify outliers. Colleagues also suggesting restricting the analysis to "standard" ganzfeld studies; however, there is little agreement about the defining features of a ganzfeld study, as highlighted in the edited discourse by Schmeidler and Edge (1999). As argued by Wiseman (2010), the null result spurred some researchers to recommend ways to create an overall significant result given their knowledge of known results. However, the Schmeidler and Edge (1999) discourse shows understanding by some researchers about the inefficacy of drawing strong conclusions from a singular analysis using known study outcomes "...risks looking like post-hoc hacking about in the data to try to rescue a disappointing result" (p. 343). Thus, the discourse amongst ganzfeld researchers around the turn of the century clearly shows recognition by some of the inherent limitations with conventional meta-analyses. But what is not answered is if researchers have adjusted their methods after this discourse.

Another core issue identified by Milton (1999) is interpreting meta-analyses that include studies of uncertain quality. Milton noted that previous meta-analyses of the ganzfeld used different quality criteria, with the number of safeguards examined in each analysis ranging from 2 to 18, making the quality scores incomparable across studies. Even with their own criteria, Milton notes that half of the previous meta-analyses scored, on average, fewer than half of the available quality points. This suggests that even when researchers create their own quality criteria, they still fail to demonstrate high levels of methodological quality. Or, they could be constructed in such a way that studies which produced null or negative results would have lower quality scores and thus have less weighting in the subsequent analysis. Either

way, Milton (1999) argued that there is a high risk earlier meta-analysis databases consist of poor-quality studies, leading to potentially inflated outcomes due to methodological flaws. Therefore, study estimates are likely the *minimum* estimates of quality due to poor reporting and incomplete capture of methodological safeguards due to the coding schemes.

In response to the null findings by Milton and Wiseman (1999), the edited volume by Schmeidler and Edge (1999) reveals that ganzfeld researchers were acutely aware of the limitations and concerns of meta-analyses at the time. One researcher notes that there is "...an opposition of philosophies: one which seeks insight and one which seeks closure" (p.346). This distinction exemplifies the earlier section on researcher degrees of freedom: a researcher with strong beliefs, post-hoc data selection and control of the decision-making process can shape their meta-analysis to achieve a desired outcome. Researchers also acknowledged the inherent flexibility and arbitrariness in conducting meta-analyses, akin to the "apples and oranges" principle (Sharpe, 1997). Recognising this issue, one researcher advocated for pre-planned sensitivity analyses, a topic still discussed to this day (e.g., Taylor & Munafò, 2016; Watt & Kennedy, 2017).

Despite the extensive discourse amongst researchers, there has not been a systematic assessment of the inclusion criteria and methods of statistical analyses within the literature. The most recent umbrella review, conducted by Tressoldi and Storm (2021a), provides a summary of meta-analysis findings only. Thus, the present systematic review uses a fresh approach to examine the psi ganzfeld debate, focusing on the methodological decisions made by the authors. The objectives are twofold: first to examine the differences in inclusion criteria among meta-analyses, which represent a form of researcher degrees of freedom; and second, to clarify how these choices, potentially influenced by researchers' beliefs, shape the objectives, results and conclusions of each analysis. Through this review, I aim to shed light on the role of researcher degrees of freedom in shaping methodological and analytical decisions. In particular, the inclusion criteria themselves – a specific concern highlighted in the Schmeidler and Edge (1999) article. The discourse emphasises the importance of identifying the factors contributing to diverging results across studies and perhaps understand the lack of consensus in the literature.

4.3 Methods

This project began as a preregistered specification-curve multiverse meta-analysis, following the analysis designed by Voracek et al. (2019). The preregistration details can be found in Publications and Contributions. The specification-curve multiverse meta-analysis is a combination of multiverse analysis (Steegen et al., 2016), as mentioned in chapter 2 and specification-curve analysis (Simonsohn et al., 2020). The methodology aims to quantitatively evaluate the influence of researcher decisions when conducting a meta-analysis. Specifically, how researchers' choices about which studies to include in their analysis (i.e., which data to analyse) and statistical models (i.e., how to analyse the data) influence final results and conclusions (Plessen et al., 2022; Voracek et al., 2019).

However, due to a misinterpretation of the methods, the analysis changed to a methodological umbrella review. By this stage, the data had been already been collected and coded by two reviewers, hence, this systematic review is not preregistered. Nonetheless, the 'How' and 'Which' factors, as used by Voracek et al. (2019), are used for this systematic review as they were applied during the initial coding of the meta-analyses. These factors were created and used to clarify the observed divergences among the meta-analyses. We expanded the scope of this review to provide a more comprehensive qualitative assessment of the psi meta-analysis literature. In addition to analysing the inclusion criteria and statistical methods, we also code the specific objectives, findings and conclusions of each meta-analysis, as per Goodyear-Smith et al. (2012). Furthermore, we assess overall study quality of each meta-analysis using the Database of Abstracts of Reviews of Effects (DARE; Centre for Reviews and Dissemination, 1995), which has a wide scope to be easily applicable across topics. The quality as assessment and objectives, findings and conclusions of each meta-analyses were conducted *after* the coding of the Which and How factors. Thus, the primary interest of this systematic review is the methodological variations within each meta-analysis of the psi ganzfeld, rather than whether the results testing the psi hypothesis are statistically significant. All screening and data extraction was conducted using Covidence systematic review software, Veritas Health Innovation.

4.3.1 Data sources and search strategy

As per the preregistration document for the quantitative analysis, the lead author conducted systematic literature searches in two bibliographic databases (PsycInfo and

Scopus) on the 20th June 2023, using the search terms “ganzfeld”, “psi”, “esp”, “meta-analysis”, “summary”, “review” and “extrasensory perception”. Reference lists were also used as a method to manually extract relevant meta-analyses during the literature search. The search results are show in the PRISMA flow chart in Figure A.1.

Inclusion criteria and screening of papers

To be included in the review, the meta-analysis had to:

- Assess the psi ganzfeld hypothesis using any design (precognition, telepathy, clairvoyance)
- If the meta-analysis assessed telepathy designs, we accepted all regardless of sender/receiver dynamic
- Use any sample (selected or unselected participants)

Studies were excluded if they were published outwith January 1980 – April 2023, did not specifically state they assessed the psi ganzfeld design, reported in a language other than English, or used a fixed-response design. As per the preregistration, metaanalyses also had to be published in peer-reviewed journals. However, during the screening and full text assessment of the meta-analyses, the lead author noticed that some unpublished and non-peer reviewed meta-analyses were cited several times. Hence, the decision was made to include non-peer reviewed meta-analyses. This also allowed the authors to assess if there were changes in the methodologies of meta-analyses if they had multiple formats, i.e., conference proceedings versus a peer-reviewed journal article.

4.3.2 Data extraction and analysis

After the database searches, the lead author screened all 488 papers. The screening resulted in a final of 21 meta-analyses (see Figure A.1). Once extracted, two authors (ALP and CW) independently extracted the number of studies, number of sessions, number of hits, hit rate, statistical outputs, effect size measures and coded the metaanalyses for Which and How factors. The Which and How factors were constructed by the same two authors, due to their prior knowledge of the psi ganzfeld literature and how the papers diverge on their inclusion and modelling methods. The factors did however change during the pilot round of coding due to nuances in the metaanalyses (detailed below). The objectives, findings and

conclusions were extracted by the lead author and verified by a research assistant. Likewise, the quality assessment was conducted by the lead author and same research assistant. We did not search any supplementary materials.

4.3.3 Objectives, Findings and Conclusions

The lead author and a research assistant extracted three characteristics from each meta-analysis, 1) the objective(s), 2) findings and, 3) conclusions, as conducted by Goodyear-Smith et al. (2012). We used this approach to better understand the given rationale for each meta-analysis and to see how these characteristics do (or do not) align with the findings of each meta-analysis.

Which and How factors

Which factors: Which data to analyse?

Regarding decisions about which groups to compare in a meta-analysis, we coded the five relevant study features:

1. *Study design.* Authors may include only one particular ganzfeld paradigm, or analyse all paradigms in their meta-analysis. We code all ganzfeld designs: 1) precognition, 2) telepathy, 3) clairvoyance, 4) all ESP designs, 5) other combination.

2. *Study mechanism.* 1. Early ganzfeld studies were manual and have since become more automated due to computer technology. Given the development of technology, there may be increased focus on auto-ganzfeld methods only, as automated procedures should reduce the likelihood of error or deliberate fraud. We code for all study mechanisms: 1) manual, 2) auto-ganzfeld, 3) mixed (partially automated), 4) not specified.
3. *Participant type.* Participants selected for certain characteristics (e.g., creativity, belief in the paranormal etc.) demonstrate higher hit rates than unselected participants[6]. A review of the psi ganzfeld literature found selected participants provide the largest

[6] Studies which use unselected participants may incidentally include participants that have these favourable characteristics; however, the term unselected participants mean there is usually no *explicit* screening of participants to ascertain their attributes.

effects and recommend sole use of selected participants in future ganzfeld studies (Baptista et al., 2015). To assess if meta-analyses have favoured including primary studies with a certain participant group, we include 1) selected participants only, 2) unselected participants only, 3) selected and unselected participants, 4) not specified.

4. *Randomisation method.* Target pool, target selection, and target presentation order has evolved from manual (e.g., rolling dice, card shuffling) to true randomisation (e.g., electronic random number generator). The manual randomisation methods are susceptible to manipulation and are not truly random. We code each meta-analysis included primary studies based on their randomisation methods. We include 1) manual randomisation only, 2) automated randomisation only, 3) any type of randomisation, 4) not specified.

5. *Inclusion of unpublished studies.*[7] Given the small field of researchers using the ganzfeld method, unpublished studies (e.g., doctoral thesis, undergraduate dissertation) are often cited in the literature. We code meta-analyses for the type of publication: 1) includes unpublished studies, 2) does not include unpublished studies and, 3) not specified.

How factors: How to analyse the data?

This factor pertains to the statistical and modelling decisions made by the authors of the meta-analysis. We diverge on this factor for two reasons. Firstly, our preregistration for the quantitative analysis differed from the approach outlined by Voracek et al. (2019) as we aimed to assess the effect sizes utilised in the literature, unlike the approach taken by Voracek et al. (2019). Secondly, our preregistration stated our intention to code the type of meta-analytic model used. However, during the coding process, we encountered difficulty as very few reports specified the type of modelling employed, leading us to abandon this aspect of coding. As our preregistration was initially intended for a quantitative analysis, this review presents the results for one How factor only, as detailed below:

1. *Effect size measure.* We code each meta-analysis for what effect size measure the authors used when conducting their statistical analysis.

[7] Thesis published on university repositories are classified as published studies, however, for this review we regard unpublished studies as any piece of research which has not gone through formal peer-review, such as an academic journal.

Two authors, ALP and CW, independently coded each meta-analysis on these factors. Coders were instructed to focus on new databases created by the authors. For example, in some meta-analyses, authors re-analysed a previous database and reported on a second, new database created based on a different set of inclusion criteria – this new database was the one of primary interest for our analysis. The coders originally followed an 'explicitly stated' paradigm, however, this was dropped after a pilot round of coding which resulted in a high proportion of Which factors being coded as 'not specified'. Thus, the coding reported is based upon a 'can be inferred' paradigm. For example, for unpublished studies, the coders assessed the reference lists to see if they included unpublished studies, a factor the authors may have not explicitly stated in their inclusion/exclusion criteria. The two authors resolved rating disputes with an in-person meeting.

4.3.4 Quality Assessment

The quality assessment was independently coded by the lead author and a research assistant, using the Bambra et al. (2009) modified version of the DARE tool (Centre for Reviews and Dissemination, 1995). After independently coding, the two coders met to resolve their disagreements. During this meeting, the two coders agreed that due to the overlap between DARE2 and DARE3, you could not have one without the other, hence the similar scores for these two criteria. Although this tool is designed for healthcare research, the quality assessment is broad enough to be applied to other literature. The adjusted DARE tool has seven criteria:

1. The presence of a well-defined research question
    a. The question should define at least the participants, the intervention, the outcomes and study design
2. The presence of a defined search strategy
    a. The research strategy should include at least one named database combined with reference checking, hand-searching, citation follow-up, or expert contact
3. The articulation of inclusion/exclusion criteria

    a. The review should make grounds for study inclusion and exclusion transparent in terms of participants, intervention, outcomes and study design 4. Whether the study design and numbers of studies are clearly stated

4. The review should outline designs of the included studies and make it clear which and how many studies are in the final synthesis
5. The presence of the description of the quality assessment process

    a. The review should clearly describe the quality assessment process, which quality appraisal tool is used and the relative quality of each included study

6. The presence of appropriate data synthesis

    a. The review should use meta-analysis or narrative synthesis, whichever is most suitable given the heterogeneity of studies and their methodological quality. If studies are very heterogeneous, narrative synthesis is appropriate

7. The involvement of at least two authors at every stage of the review process

    a. To minimise bias, the review should have at least two reviewers involved in each stage (study selection, data extraction, quality appraisal, synthesis) of the review

Each meta-analysis is given a point for the presence of each criterion, with greater number of points indicating a higher quality meta-analysis. As our quality assessment was conducted after the screening and data extraction of the meta-analyses and we wanted to assess all relevant papers, we did not use the quality assessment as an include/exclude tool, as it is commonly used (see Bambra et al., 2009).

4.4 Results

As shown in Table 4.1, the primary objective of most meta-analyse is to evaluate the existence of anomalous cognition (i.e., the psi hypothesis) using the psi ganzfeld method. However, certain meta-analyses addressed more nuance aspects of the literature, such as the type of statistical analysis (Milton, 1997a; Pooley et al., 2023) or internal factors such as participant characteristics, participant pairing and software design (Honorton et al., 1990). The majority of meta-analyses are published in academic journals, and the critical responses

to them also tend to remain within the academic sphere. However, two meta-analyses published in popular science books (Radin, 2006; Schlitz & Radin, 2003) report overwhelmingly supportive evidence for psi using the ganzfeld method.

Study selection and inclusion criteria often evoke a debate between researchers given the divergence across meta-analyses. Exemplary debates among authors include Hyman (1985) and Honorton (1985), who examined the same database created by Honorton and came to contrasting findings, as discussed in chapter 1. Another notable debate arose from the Milton and Wiseman (1999) paper, which sparked extensive discussion within the field due to their non-significant findings of studies published between 1987 and February 1997. This led to a response from Bem et al. (2001) and Storm and Ertel (2001), with the former reporting supportive evidence, after including 10 studies published *after* the Milton (1999) cut-off date. Storm and Ertel (2001) heavily criticised the Milton and Wiseman (1999) analysis on several grounds, including "spurious judgement calls", such as excluding studies predating 1986 (before the *Joint Communiqué*; Hyman and Honorton (1986). They reported statistically significant results upon re-analysing the database, incorporating precut-off studies and combining independent databases. More recently, Storm et al. (2010b) published positive results and claimed further replication of the psi hypothesis. However, they received criticism from Rouder et al. (2013), who questioned why Storm and colleagues did not include negative studies within two publications (Del Prete & Tressoldi, 2005; Tressoldi & Del Prete, 2007) given their own inclusion criteria. Rouder et al. (2013, p. 243) declared "these two omissions are an example of selection artifact", perhaps indicating a selection bias, as argued by Goodyear-Smith et al. (2012). Notably, these positive studies from the same two papers were included in the Storm et al. (2010b) meta-analysis. Another point of contention between the two meta-analyses is the inclusion of another study, (May, 2007). When Rouder et al. (2013) re-analysed the Storm et al. (2010b) database using a Bayesian framework, they excluded May (2007) as it had hard to interpret results and lacked internal validity. Their Bayesian analysis found overwhelming evidence against the psi hypothesis. Storm et al. (2013) disputed Rouder et al. (2013) "dubious justification" for excluding studies and asserted that the May (2007) study was appropriate for inclusion. It is noteworthy that the May (2007) study reports highly significant findings (64% hit rate). Further, when Storm et al. (2013) adopted a Bayesian framework, like Rouder et al. (2013), they once again found positive replication of the psi hypothesis, similar to their original 2010 meta-analysis.

These differences in study inclusion and selection criteria may reflect the prior beliefs of the researchers, manifesting through selection artefact, HARKing, and other related issues. Nonetheless, the psi ganzfeld literature underscores the inherent flexibility and ambiguity encountered by all researchers when constructing and conducting a meta-analysis. These inclusion decisions are examined in detail below.

4.4.1 Objectives, findings, conclusions

The objectives of the psi ganzfeld meta-analyses (see Table 4.1) can be broadly categorised into three main groups: 1) those that investigate the evidence of psi using the ganzfeld methodology, 2) those that concentrate on internal ganzfeld study design features and, 3) those that address methodological and/or statistical issues.

Among the meta-analyses reviewed, the majority (11 out of 22) fall into the first group (Bem & Honorton, 1994; Honorton, 1985, 1985; Milton & Wiseman, 1999; Radin, 2006; Schlitz & Radin, 2003; Storm et al., 2010b; Storm & Ertel, 2001; Storm & Tressoldi, 2020; Tressoldi & Storm, 2023; Williams, 2011) Four meta-analyses focus on internal factors, such as the novel autoganzfeld design (Honorton et al., 1990), the presence of a sender (Honorton, 1995), the relationship between extraversion and study performance (Honorton et al., 1998), and the influence of five telepathy design features on study outcome (Pooley et al., 2023). In the third group, six meta-analyses (out of 22) focus on methodological and/or statistical issues, including the sensitivity of direct hits versus ranks (Milton, 1997a) the creation of a "less biased" study standardness quality assessment (Bem et al., 2001; Palmer & Broughton, 2000), and the appropriateness of Bayesian statistics for analysing psi ganzfeld literature (Rouder et al., 2013; Storm et al., 2013; Utts et al., 2010).

Unsurprisingly, the findings and conclusions of each meta-analysis align with their respective objectives. However, determining which came first is challenging. Most meta-analyses report positive findings, suggestive sufficient evidence to support the existence of a replicable effect of psi in the ganzfeld (Bem & Honorton, 1994; Bem et al., 2001; Honorton, 1985; Storm & Ertel, 2001; Storm & Tressoldi, 2020; Storm et al., 2010b; Tressoldi & Storm, 2023; Williams, 2011). Whereas, others are more cautious, citing methodological concerns that hinder confirmation of anomalous cognition (Hyman, 1985; Milton & Wiseman, 1999; Rouder et al., 2013). Nonetheless, discourse among authors with diverging results often exhibit mutual support. For instance, the discourse between Hyman (1985) and Honorton (1985) resulted in the *Joint Communiqué*, establishing methodological and statistical

improvements (Hyman & Honorton, 1986). Similarly, Rouder et al. (2013) concurred with Storm et al. (2010b) that uncritical consideration of recent psi experiments could strongly support the psi hypothesis, but emphasised the need for more critical and detailed assessment. Furthermore, Storm et al. (2013) acknowledged the suggestion to use Bayesian statistics to assess the literature and implemented this approach, despite arriving at different conclusions than Rouder et al. (2013).

Table 4.1
Objectives, Findings and Conclusions for each Meta-Analysis, Sorted by Year of Publication

Meta-analysis	Objectives	Findings	Conclusions
Honorton (1985)	**Assess if a statistically significant effect in psi ganzfeld database; impact of reporting bias; relationship between study flaws and outcomes**	**Cumulative *z*-scores show probability not larger than one part in 100 million; 43% of studies independently significant**	**Significant psi ganzfeld effect; no relationship between study outcomes and study quality**
Hyman (1985)	Critical assessment of field of parapsychology; program which consists of studies by a variety of investigators	Inadequate randomisation, insufficient documentation correlate with study significance and ES; inadequate security and sensory leakage	Bias in reporting of results; multiple testing; procedural flaws; statistical errors; too many problems for evidence
Honorton et al. (1990)	**Assess evidence for psi in autoganzfeld; dynamic vs static targets; effects of sender/receiver relationship; compare with manual ganzfeld**	**Significant effect; homogeneous; dynamic targets higher success than static; PRL results similar to those in previous meta-analyses**	**Cumulative evidence to conclude ganzfeld represents genuine communication anomaly**
Bem & Honorton (1994)	Present a meta-analysis to wider psychology community	Highly significant result; medium ES; dynamic targets higher success than static	Stringent standards from *Joint Communiqué* are met; reliable relationship between psi performance and certain experimental and subject variables
Honorton (1995)	**Evaluate magnitude of effect overall and as function of target presentation conditions**	**Combined evidence provides strong evidence for anomalous cognition in ganzfeld**	**Fail-safe estimates render alternative explanations untenable**
Milton (1997)	Determine whether direct hits or ranks are more sensitive	Sum of ranks outperforms direct hits for ES and *p*-values but not significantly different	Researchers are freer than they may think to use sum of ranks
Honorton et al. (1998)	**Examine the relationship between psi performance and extraversion**	**Free-response individual testing significant and homogeneous; significant relationship between extraversion and psi task**	**Support for relationship between extraversion and free-response ESP task; not affected by artefacts or validity issues**

Milton & Wiseman (1999)	Review of new ganzfeld studies to assess if a broader range of investigators successfully replicated autoganzfeld	Non-significant main effect; new studies did not confirm main effect of autoganzfeld studies; 1/3 internal effect replicated	New ganzfeld studies show near-zero ES; statistically non-significant overall cumulation
Palmer & Broughton (2000)	**Create a less biased evaluation of standardness of ESP studies**	**10 studies published after M&W show marked improvement in outcomes; difference not significant**	**Manual ganzfeld and PRL heterogeneous; methodological standardness significantly correlated with ES**
Bem et al. (2001)	Test for methodological standardness due to decline in ganzfeld ES	10 new ganzfeld replication studies statistically significant	Ganzfeld remains replicable; higher protocol standardness positively and significantly correlated with ES
Storm & Ertel (2001)	**Replace M&W meta-analysis based on unified ganzfeld data; include studies overlooked by M&W**	**Manual ganzfeld studies prior to *JC* highly significant; quality-rated ES non-significant**	**Psi ganzfeld replicable technique for producing psi effects in lab; B&H findings replicated**
Schlitz & Radin (2003)	N/A	N/A	Results far beyond chance expectation; high levels of consistency across 10 independent labs; unlikely systematic methodological flaws
Radin (2006)	**N/A**	**Combined hit rate 32%**	**Select reporting not an issue for these studies**
Utts et al. (2010)	Demonstrate Bayesian analysis, prior belief and knowledge can be incorporated into analysis	N/A	Psi studies natural context for Bayesian analysis
Storm et al. (2010b)	**Produce a comprehensive, up-to-date meta-analysis of ganzfeld and other free-response noise-reducing paradigms**	**Ganzfeld significantly greater than other free-response designs; selected participants more successful than unselected, only in ganzfeld**	**Consistency of results demonstrated in data; replication by range of investigators**
Williams (2011)	Assess post-*JC* replication status of ganzfeld database; update and confirm results of previous meta-analyses	Consistent findings for B&H, M&W, Storm (2010b) databases; 10 new studies in Bem (2001) non-significant but 40 combined is	Findings consistent with earlier meta-analyses; significant overall hit-rate in studies post-*JC;* finding remains apart from early ganzfeld and PRL
Rouder et al. (2013)	**Bayes factor assessment for Storm (2010b); Bayes factor informs reader about beliefs**	**Manual randomisation produces better psi performance than computer randomised; not psi**	**Randomisation affects outcome; file-drawer; selectivity artefacts**

Storm et al. (2013)	Re-assess database due to critique about study inclusion and statistics	Findings upheld with Bayesian approach	Analysis of reduced Storm (2010b) database using Bayesian has same findings as original
Storm & Tressoldi (2020)	**Test ESP in free-response studies 2009-2018**	**9 new ganzfeld studies homogeneous; support for ESP in ganzfeld**	**Findings replicate earlier database results**
Pooley et al. (2023)	How 5 telepathy design features influence study outcome; different ES not based on *z*-scores	Review period decreases study success; model with all studies is heterogeneous; evidence of publication bias	Homogenous database shows ganzfeld produces significantly greater than chance when using different ES
Tressoldi & Storm (2023)	**Investigate anomalous perception; moderators that affect task performance**	**Frequentist and Bayesian random-effect models reject null hypothesis with high probability**	**Sufficient evidence for anomalous perception in ganzfeld; no publication bias; no QRPs**

Note. Abbreviations have been used to improve table formatting, MA: meta-analysis; M&W: Milton and Wiseman (1999); B&H: Bem and Honorton (1994); JC: Joint Communiqué, Hyman and Honorton (1986); ES: Effect size.

4.4.2 Which and How factors

The summary of the inclusion criteria, including the Which and How factors for each meta-analysis are shown in Table 4.2.

Study design

This Which factor focuses on which ganzfeld design researchers included in their meta-analysis. As shown in Table 4.2, the majority of meta-analysis looking at the psi ganzfeld include all types of designs (precognition, telepathy, clairvoyance). However, older meta-analyses tend to favour telepathy and combinations of two designs. This reflects the active research area of the time with telepathy and clairvoyance designs being favoured in primary research. As more studies have been conducted, the meta-analyses appear to reflect this change by including all designs.

Study mechanism

The psi ganzfeld studies evolved from solely manual to automated (autoganzfeld) methods from the late 1980s. This is reflected in the meta-analysis inclusion criteria; older meta-analyses focused solely on manual methods but the change is demonstrated by Honorton et al. (1990) and Bem and Honorton (1994) which only include autoganzfeld studies. These two meta-analyses assessed whether the automated studies replicate the effects reported in the manual studies, which both papers suggest they do. Over time, meta-analyses from the early 2000s onward began including both manual and autoganzfeld, even though the homogeneity of these databases is contested (Hyman, 2010; Milton & Wiseman, 1999). However, the researchers within the field are aware that older, manual ganzfeld studies produce greater hit rates and effect sizes than the automated studies (Milton & Wiseman, 1999; Storm et al., 2010b). The standout meta-analysis is Storm and Tressoldi (2020), which only includes autoganzfeld studies. This paper was coded as such given the year range of primary studies included (2009-2018) as the authors updated their 2010b analysis.

Participant type

As noted earlier, it is worth exploring whether meta-analyses in this field tend to favour primary studies than only used selected participants, as recommended by (Baptista et al., 2015). However, all meta-analyses coded include both unselected and selected participants. Yet, within these meta-analyses, very few directly assess the type of participant on the task within their meta-analytic models (see Honorton et al., 1990; Honorton, 1997; Honorton et

al., 1998). Milton and Wiseman (1999) assessed the internal effects reported by Bem and Honorton (1994) but only one internal effect was replicated. More recently, Storm et al. (2010b) found an interaction between free-response study design and participant type. This finding was replicated in their update (Storm & Tressoldi, 2020).

Randomisation method

Concerns regarding target randomisation have been a consistent point of contention across psi ganzfeld meta-analyses since the 1980s (see Honorton, 1985). This is shown in Table 4.2, where meta-analyses actively select for automated randomisation or both types, with none including solely manually randomised studies. Example debates between meta-analyses include, Rouder et al. (2013), who reported a significant effect of randomisation type on study outcome, indicating that older, manually randomised ganzfeld studies are more successful, while newer, automatically randomised reduce the likelihood of overall study success. However, this finding was contested by Storm et al. (2013), who found the opposite. Likewise, Storm and Ertel (2001) found manual ganzfeld studies conducted prior to the *Joint Communiqué* (Hyman & Honorton, 1986) were highly significant, but not after they were quality weighted, unlike Hyman (1985).

Inclusion of unpublished studies

Finally, the inclusion (or exclusion) of unpublished studies is a pertinent question, particularly considering the file-drawer discourse associated with parapsychological research. Notably, meta-analyses published between 1997-2001 did not include unpublished studies. This could be attributed to two potential reasons: first, the researchers might have genuinely believed there was no file-drawer problem within the literature and consequently did not actively seek unpublished studies. Second, if a file-drawer problem does indeed exist, not searching for unpublished studies might favour authors who are more open to psi, as studies with non-significant results are less likely to be published. This trend is interesting, particularly since most researchers publishing during this period reported favourable conclusions (with the exception of Milton and Wiseman, 1999). Since 2001, peer-reviewed meta-analyses have more varied inclusion of unpublished studies. Interestingly, the Storm and Tressoldi (2020) update to their 2010b paper does not include unpublished studies, unlike the first analysis. Perhaps either due to a change in their own selection criteria they did not explicitly state or due to no relevant studies in unpublished formats met the inclusion criteria.

Effect size measure

The most commonly used effect size estimate in the psi ganzfeld meta-analysis $_q$ [8] with 10 assessed meta-analyses using this formula. Earlier literature is $z/\sqrt{n}$ meta-analyses used a variety of effect size measures, suggesting a calibration period of how to assess the ganzfeld statistics given the use of binomial hits and binomial *z*-scores. Alternative effect size measures have been used to assess how they may also impact the conclusions of each meta-analysis, such as Milton (1997b) comparing hits and ranks. Likewise, Rouder et al. (2013) used a Bayes factor analysis to ascertain if the positive findings of Storm et al. (2010b) withheld a Bayesian framework. More recently, Pooley et al. (2023) used two alternative effect size estimates due to the concerns with the commonly used effect size within the literature.

[8] Note this effect size is the same as $z/N^{1/2}$ which is a common alternative description in the ganzfeld meta-analysis.

Table 4.2
Summary of Study Which and How Factors for each Meta-Analysis, Sorted by Year of Publication

Meta-analysis	Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies	Effect size	*ES* value
Honorton (1985)	**Combination**	**Manual**	**Both**	**Any**	**Yes**	**Stouffer's *Z***	**6.60**
Hyman (1985)	All	Manual	Both	Any	Yes	Freeman-Tukey arc sine (binomial)	5.98
Honorton et al. (1990)	**Combination**	**Autoganzfeld**	**Both**	**Automated**	**Yes**	**Cohen's *h***	**0.2**
Bem & Honorton (1994)	Telepathy	Autoganzfeld	Both	Automated	No	Pi	0.59
Honorton (1995)	**Combination**	**Mixed**	**Both**	**Any**	**Yes**	**Cohen's *h***	**0.16**
Milton (1997)	Combination	Mixed	Both	Any	No	$z/\sqrt{(n)}$ & Stouffer's Z	2.49 (Stouffer's Z)
Honorton et al. (1998)	**Combination**	**Autoganzfeld**	**Both**	**Automated**	**No**	***r***	
Milton & Wiseman (1999)	All	Mixed	Both	Any	No	$z/\sqrt{(n)}$	0.013
Palmer & Broughton (2000)	**All**	**Autoganzfeld**	**Both**	**Automated**	**No**	$\boldsymbol{z}/\sqrt{(\boldsymbol{n})}$	**0.165**
Bem et al. (2001)	Combination	Autoganzfeld	Both	Automated	No	$z/\sqrt{(n)}$	0.17
Storm & Ertel (2001)[a]	**All**	**Manual**	**Both**	**Any**	**No**	$\boldsymbol{z}/\sqrt{(\boldsymbol{n})}$	**0.222**
Schlitz & Radin (2003)	All	Mixed	Both	Any	Not specified		
Radin (2006)	**All**	**Mixed**	**Both**	**Any**	**Not specified**	$\boldsymbol{z}/\sqrt{(\boldsymbol{n})}$	**0.16**

Utts et al. (2010)	Combination	Mixed	Both	Any	Not specified	Bayesian	
Storm et al. (2010b)[b]	**All**	**Mixed**	**Both**	**Any**	**Yes**	**$z/\sqrt{(n)}$ & Stouffer's Z**	**0.152 6.34**
Williams (2011)	All	Mixed	Both	Any	Yes	N/A	N/A
Rouder et al. (2013)[c]	**All**	**Mixed**	**Both**	**Automated**	**Yes**	**Bayes factor**	**63.3 to 1**
Storm et al. (2013)	All	Mixed	Both	Any	Yes	$z/\sqrt{(n)}$ & Bayes factor	0.14
Storm & Tressoldi (2020)	**All**	**Autoganzfeld**	**Both**	**Automated**	**No**	**$z/\sqrt{(n)}$**	**0.119**
Pooley et al. (2023)[d]	Telepathy	Mixed	Both	Any	Yes	Binomial hits as mean & binomial hits as proportion & Stouffer's Z	5.81 (Stouffer's Z)
Tressoldi & Storm (2023)	**All**	**Mixed**	**Both**	**Any**	**No**	**$z/\sqrt{(n)}$**	**.099**

[a] non-quality rated outcomes reported [b] homogeneous ganzfeld dataset [c] Revised Set 1 ganzfeld only [d] non-homogeneous dataset

4.4.3 Quality assessment

Table 4.3 summarises the quality assessment for each meta-analysis, using the adjusted DARE guidelines. The average quality score for the ganzfeld meta-analyses is 3.7 (range 0-6) with none of the assessed articles receiving the full score of 7. However, the average increases to 4.4. when excluding meta-analyses not published in academic journals, including the two popular science books (Radin, 2006; Schlitz & Radin, 2003), the workshop tool (Utts et al., 2010) and the review by Williams (2011). The involvement of at least two authors at every stage of the review process (DARE7) was the most commonly unmet criteria among articles. The only article to receive a point for this was Tressoldi and Storm (2023). However, I acknowledge that each meta-analysis may have indeed had at least two researchers at every stage of the review process, but this was not made explicitly clear nor was it deducible for each analysis. Given pre-registration practices and increase reporting of all stages of reports, such as supplementary materials and online repositories, this criterion is likely to be increasingly met. However, for older meta-analyses, this cannot be assumed.

In terms of reproducibility, transparency and researcher degrees of freedom, the first four DARE criteria are of the most interest. The vast majority of meta-analyses demonstrate the presence of a well-defined research question (DARE1). Notably, those that do not meet this criterion include the popular science books, the workshop paper, the review by Williams (2011) and, interestingly the Bem and Honorton (1994) article. For DARE2 and 3, approximately half of the assessed meta-analyses have defined search strategies and clear inclusion/exclusion criteria. One paper that deviates from the scoring agreed-upon by the two coders is Storm and Ertel (2001). While the article provides very detailed inclusion/exclusion criteria, it does not define the journals searched or the search strategy used to create their database. However, the remaining assessed meta-analyses align with the agreed-upon coding. For DARE4, the majority of reviews clearly state which studies they include in their analysis, usually presented in tables and/or indicated in the reference list. Again, meta-analyses in the two popular science books do not specify details.

It must be noted that the articles in the Storm-Rouder discourse (Rouder et al., 2013; Storm et al., 2013) exhibit notably low compliance across all DARE criteria, especially for the second, third and fourth DARE. Since they are both re-analysing the same database as Storm et al. (2010b), albeit with slight alterations, they do not provide inclusion/exclusion criteria. These criteria could be inferred, but as noted in the Discussion section, each meta-

analysis was taken at face value. Another analysis which is low across all criteria is the review by Williams (2011). The authors states in the article it is not a formal meta-analysis, but as this article is frequently cited in the psi ganzfeld literature as one, it is included in this review.

The assessment of study quality, represented by DARE5, is notably lacking in the psi ganzfeld meta-analysis literature. Due to the unique nature of the experiment, there is no universally agreed-upon standard for assessing study quality. Instead, researchers often devise their own criteria (e.g., Honorton, 1985; Hyman, 1985; Storm & Ertel, 2001) or utilise combinations or previously used quality scales within the parapsychology literature (e.g., Storm & Tressoldi, 2020; Storm et al., 2010b).

Table 4.3
Quality Criteria for each Meta-Analysis using the DARE criterion, Sorted by Year of Publication

Meta-analysis	DARE1	DARE2	DARE3	DARE4	DARE5	DARE6	DARE7	Total	*ES* measure	*ES*
Honorton (1985)	**YES**	**NO**	**NO**	**YES**	**YES**	**YES**	**NO**	**3**	**Stouffer's *Z***	**6.60**
Hyman (1985)	YES	YES	YES	YES	YES	YES	NO	6	Freeman-Tukey arc sine (binomial)	5.98
Honorton et al. (1990)	**YES**	**NO**	**NO**	**YES**	**NO**	**YES**	**NO**	**3**	**Cohen's *h***	**0.2**
Bem & Honorton (1994)	NO	NO	NO	YES	NO	YES	NO	2	Pi	0.59
Honorton (1995)	**YES**	**YES**	**YES**	**YES**	**NO**	**YES**	**NO**	**5**	**Cohen's *h***	**0.16**
Milton (1997)	YES	YES	YES	YES	NO	YES	NO	5	$^{z}/_{\sqrt{(n)}}$ & Stouffer's Z	0.07 2.49
Honorton et al. (1998)	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**NO**	**6**	***r***	
Milton & Wiseman (1999)	YES	NO	NO	YES	NO	YES	NO	3	$^{z}/_{\sqrt{(n)}}$	0.013
Palmer & Broughton (2000)	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**NO**	**6**	$^{\boldsymbol{z}}/_{\sqrt{(\boldsymbol{n})}}$	**0.165**
Bem et al. (2001)	YES	NO	NO	YES	YES	YES	NO	4	$^{z}/_{\sqrt{(n)}}$	0.17
Storm & Ertel (2001)	**YES**	**NO**	**YES**	**YES**	**YES**	**YES**	**NO**	**5**	$^{\boldsymbol{z}}/_{\sqrt{(\boldsymbol{n})}}$	**0.222**
Schlitz & Radin (2003)	NO	NO	NO	NO	NO	NO	NO	0		
Radin (2006)	**NO**	**NO**	**NO**	**NO**	**NO**	**YES**	**NO**	**1**	$^{\boldsymbol{z}}/_{\sqrt{(\boldsymbol{n})}}$	**0.16**

Utts et al. (2010)	NO	NO	NO	YES	NO	YES	NO	2	Bayesian	
Storm et al. (2010b)	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**NO**	**6**	**$z/\sqrt{(n)}$ & Stouffer's Z**	**0.152 6.34**
Williams (2011)	NO	NO	NO	YES	NO	NO	NO	1	N/A	N/A
Rouder et al. (2013)	**YES**	**NO**	**NO**	**NO**	**NO**	**YES**	**NO**	**2**	**Bayes factor**	**63.3 to 1**
Storm et al. (2013)	YES	NO	NO	NO	NO	YES	NO	1	$z/\sqrt{(n)}$ & Bayes factor	0.14
Storm & Tressoldi (2020)	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**NO**	**6**	**$z/\sqrt{(n)}$**	**0.119**
Pooley et al. (2023)	YES	YES	YES	YES	NO	YES	NO	5	Binomial hits as mean & binomial hits as proportion & Stouffer's Z	5.81 (Stouffer's Z)
Tressoldi & Storm (2023)	**YES**	**YES**	**YES**	**YES**	**NO**	**YES**	**YES**	**6**	**$z/\sqrt{(n)}$**	**.099**
Yes (%)	75	45	50	85	35	95	5			
No (%)	25	55	50	15	65	5	97			

4.5 Discussion

The analysis of 21 psi ganzfeld meta-analyses uncovered several novel insights into the methodological approaches, challenges, and debates within this field. Firstly, our examination of the objectives for each meta-analysis sheds light on the multifaceted nature of research aims within the psi ganzfeld literature. While many focused on assessing the overall evidence for psi using the ganzfeld design, others focused on specific aspects of the study design, methodological issues, or statistical considerations. Second, our examination of the inclusion criteria (Which and How factors) revealed notable divergence among meta-analyses, reflecting varying perspectives on what constitutes relevant evidence for psi phenomena. This highlights the nuanced nature of study selection and the influence of researcher degrees of freedom on meta-analytic outcomes. Furthermore, our evaluation of study quality criteria underscored the lack of consensus in this area, with researchers employing disparate approaches to assess the quality of included studies. The lack of consistency in inclusion criteria and study quality assessments points to a fundamental challenge within the field, complicating comparisons across meta-analyses and raising questions about the reliability and validity of findings. Indeed, the absence of a single or refined set of quality criteria has been a topic of discussion within the ganzfeld discourse for some time. Some researchers argue that:

> "...quality scales are arbitrary – most of the analytical conclusions drawn themselves are more deeply flawed than the meta-analysis about which they allegedly inform us" (Schmeidler & Edge, 1999, p. 354).

One example is the criteria created by Bem et al. (2001). Their quality criteria aimed to mitigate bias by assigning higher ratings to studies that included creative participants, under the premise that such inclusion was indicative of methodological rigour. However, by this time point, only a limited number of psi ganzfeld studies explicitly involved creative participants. This raises questions about the underlying motivations behind such criteria (Wiseman, 2010). It raises questions as to whether it reflects conscious effort to elevate the significance of highly positive studies versus merely a reflection of current trends in participant selection within the ganzfeld paradigm. Regardless, this instance highlights the considerable freedom wielded by researchers in determining study inclusion, devising quality criteria, and selecting statistical models – a flexibility that underscores the inherent variability and subjectivity inherent in systematic reviews. Another example is the quality scales used by

Storm and Tressoldi (2020) differing from those in their previous analysis (Storm et al., 2010b), with no justification provided for the change. The authors have every right to use a different quality scale, but direct comparisons and replication of findings are not as straightforward. Thus, the number of quality scales demonstrates the concern raised by Milton (1999) and how these cannot be compared and likely fail to capture the full methodological rigour of each study. Moreover, the researcher degrees of freedom are highlighted in the Storm and Tressoldi (2020) analysis, which includes a study that should not have been part of their analysis (Parra & Argibay, 2007). While the authors justify this decision by citing oversight in their previous meta-analysis (Storm et al., 2010b), which covered studies published between 1997-2008, they instead include the study in their update which, by their own inclusion criteria, should only have studies published between 2009-2018.

Although this review is qualitative in nature, we can make preliminary observations by comparing study inclusion and quality criteria against reported effect sizes (see Tables 4.2 and 4.3). Some existing literature states that meta-analyses including manual ganzfeld study mechanisms are more likely to report inflated hit rates (e.g., Milton & Wiseman, 1999; Storm et al., 2010b). However, this trend is not consistently found across all studies examined here. For instance, both Palmer and Broughton (2000) and Tressoldi and Storm (2023) used the same effect size measure and had the same level of quality (6/7 DARE criteria). Palmer and Broughton included only autoganzfeld studies and those with automated randomisation, while Tressoldi and Storm included all study types and any form of randomisation. Yet, Palmer and Broughton reported a higher effect size ($ES = 0.165$) compared to Tressoldi and Storm ($ES = 0.099$), contrary to expectations based on prevailing claims in the literature. That said, the highest effect size (0.22) observed using the same *ES* measure was reported by Storm and Ertel (2001), who included only manual studies and any form of randomisation, aligning with the notion that manual procedures may contribute to inflated outcomes. These mixed findings suggest that the inclusion criteria and quality of the meta-analytic process jointly influence the final effect size estimates. While this review has qualitatively assessed these dimensions, future quantitative analysis examining the relationship between inclusion criteria and study quality will offer more precise insight into how these factors contributed to variability in meta-analytic outcomes.

Our review demonstrates that the psi ganzfeld meta-analysis literature remains susceptible to allegations of cherry-picking and the tendency to explain away null results

(Wiseman, 2010). Even seemingly minor subjective decisions can accumulate, leading to divergent findings and introducing more complexity than clarity into the discourse (Schmeidler & Edge, 1999). These subjective decisions, whether consciously or unconsciously driven by prior beliefs, have the potential to significantly influence the outcomes of systematic reviews (Goodyear-Smith et al., 2012). Furthermore, the proliferation of meta-analyses, coupled with vested interests and questionable selection choices, raises concerns about the objectivity and reliability of these analyses (Ioannidis, 2016). Just focusing on the nine studies which address the psi hypothesis, the Which factors for these studies vary on all factors, apart from participant type. Thus, readers evaluating a meta-analysis are confronted with two fundamental questions: who is right, and what factors influenced the researchers' decisions? (Goodyear-Smith et al., 2012).

We acknowledge Baptista and Derakhshani (2014)'s distinction between publication and selection bias. Publication bias, inherent due to search strategy, often results in psychology publications leaning less favourably towards non-significant studies. Whereas, selection bias stems from inclusion criteria guiding the inclusion or disregard of certain studies. While Baptista and Derarkshani argue that selection bias is untenable within parapsychology, due to the early emergence of the file drawer debate, the influence of prior knowledge of results when designing a meta-analysis cannot be ignored, especially in a small field. As highlighted by Schmeidler and Edge (1999), inclusion criteria can determine the outcome of a meta-analysis "according to the taste of the analyst" (p. 337), potentially leading to post-hoc data selection, where any combination of studies and selection criteria can be applied without justification.

4.5.1 Limitations

As this systematic review is an illustrative examination of the manifestation of researcher degrees of freedom in a defined research area, rather than comprehensively including all possible databases, we constrained our analyses in several respects. First, if there were multiple databases reported in one article, we only focused on the first reported. Second, supplementary materials were not checked due to time constraints, potentially missing other information not defined in the main analysis description. Constraints such as word count limits in journals and conference proceedings may have let to the omission of crucial information that could have aided the coding process, especially the Which and How factors. Additional limitations include the evaluation of the objectives, findings and

conclusions may help better understand the rationale for each meta-analysis, but it does not lessen the potential issue of HARKing; authors could have easily changed their objectives after knowing the results.

Furthermore, our coding approach for the Which and How factors, while strongly inferred, can be interpreted differently depending on the coder. In particular, factors such as randomisation procedures and participant selection. For instance, Storm et al. (2010b) specified inclusion criteria limited to number randomisation or random-number table procedures. However, Rouder et al. (2013) noted that Dalton (1997) did not detail how randomisation was conducted. When Rouder adjusted Storm et al. (2010b) database to exclude studies using manual randomisation, contrary to Storm's criteria, they observed a significant shift in conclusions from overwhelmingly supporting psi to one which was far from favourable. Yet, in response, Storm et al. (2013) conducted a Bayesian analysis, like Rouder, including studies using random number tables and once again found favourable evidence. Similarly for participant type, Baptista et al. (2015) argued that the Milton and Wiseman (1999) database may have included selected participants, although this was not stated (or implied) in the original analysis. Likewise, the Psychophysical Research Laboratory (PRL) database (Honorton et al., 1990) did not formally select participants, yet likely had a high proportion of favourable traits as they sought to avoid university student samples (Baptista et al., 2015). Thus, this raises the question of which account the reader should trust: if one were to replicate a meta-analysis, they could only work with the latest database and results. Therefore, during the initial coding phase, we opted to take each article at face value, disregarding any subsequent clarifications. Likewise, subjectivity arose when interpreting factors with limited information, such as the design of included studies or publication type included. For instance, a study may seem to exclusively include published studies, but this may be due to no relevant unpublished studies meeting the inclusion criteria, rather than deliberate exclusion by the authors.

Finally, there are two fundamental issues which extend beyond our systematic review. First, our own method is liable to subjectivity and researcher degrees of freedom, such as the use and construction of the Which and How factors and the quality assessment measure. If other researchers conducted a methodological review of the same literature, they would likely have different conclusions. Thus, we emphasise this an exploratory project and one way to quantify and understand the psi ganzfeld literature. Second, there is no agreed upon definition of the psi ganzfeld (Schmeidler & Edge, 1999). Given the varied definitions used across

meta-analyses and the absence of a universally agreed-upon definition, establishing consensus on the core attributes of the ganzfeld is essential before claiming replicable evidence of psi phenomena through meta-analysis. If there is no agreement on the psi ganzfeld paradigm, the primary level studies themselves may be too dissimilar to aggregate effectively for meta-analysis.

4.5.2 Recommendations

Given the potential influence of prior belief when conducting meta-analyses, explicit assessment is warranted. Panagiotou and Ioannidis (2012) found that researchers' interpretation of meta-analysis results is influenced by their involvement in the field and their own findings. A similar investigation within parapsychology could shed light on the controversy surrounding the studied phenomena. Recent research suggests that psi researchers share cognitive styles with sceptics, indicating a commitment to evidence-based reasoning (Pehlivanova et al., 2024). For contentious topics and research which currently lacks consensus, it would be valuable for researchers conducting a meta-analysis to state their priors about what results they expect to find so belief can be mapped to results. This is something general psychology can adopt from parapsychologists, such as the *Journal of Anomalous Experiences and Cognition* asking researchers to declare their belief about the psi hypothesis, for example. Likewise, journal editors can ask meta-analysts to explicitly justify the need for conducting another review, to lessen the amount of redundant and overlapping meta-analyses (Sharpe & Poets, 2020).

Further, it is unwise to consider a single meta-analysis as definitive. Instead, we should embrace triangulation through different synthesis methods (Taylor & Munafò, 2016). Quantitative assessments of methodological and statistical decisions, such as specification-curve multiverse meta-analysis (Plessen et al., 2022; Voracek et al., 2019), have not been explored in the psi ganzfeld and other controversial topics. Conducting such an analysis with the psi ganzfeld literature can provide additional perspective to the discourse. This methodological umbrella review can act as a template for a specification-curve multiverse meta-analysis and code the primary level studies using the Which factors used here. By conducting a qualitative and quantitative assessment of meta-analyses, it can provide birds-eye view of an entire body of literature that is less influenced by flexibility in study selection (Plessen et al., 2022).

Finally, while adversarial collaboration is gaining traction in psychology (Clark & Tetlock, 2023; Cowan et al., 2020), controversial fields, such as parapsychology, present an opportunity to resolve persistent debates between opposing camps by fostering active collaboration. Indeed, one example of this is the open peer-review of an ongoing meta-analysis by Tressoldi and Storm (2023). But, other fields within psychology could benefit, such as non-definitive meta-analyses assessing the effect of violent video games on aggression (see Voracek et al., 2019).

Another recommendation is registration-based prospective meta-analysis (see Watt & Kennedy, 2017). This will lessen researcher degrees of freedom by preregistering methodological decisions, such as the inclusion/exclusion criteria before the results of the included studies are known (see also Sharpe & Poets, 2020). More so, these primary level studies themselves should also be pre-registered in peer-reviewed repositories to reduce bias in the individual study and subsequent meta-analysis (Watt & Kennedy, 2017). Finally, when stating hypotheses (pre-registered or otherwise), clarifying if they are exploratory or confirmatory is necessary. Exploratory research is the starting point for research, but confirmatory research is the convincing evidence that makes science valid and self-correcting (Kennedy, 2016).

4.6 Conclusion

The psi ganzfeld meta-analysis literature exemplifies the inherent variability and influence researchers have when conducting a meta-analysis. While there is no 'right' or 'wrong' way to conduct a meta-analysis, the flexibility in study design and analytical decisions introduces substantial researcher degrees of freedom, which in turn creates a susceptibility to questionable research practices—intentional or otherwise. This is particularly relevant in the context of psi research, which prides itself on methodological improvements, including pioneering registered reports, discussing the file-drawer problem and promoting greater transparency (Wiseman et al., 2019). Yet, it is not immune from the influence of researcher degrees of freedom and often comes under scrutiny by parapsychologists and non-parapsychologists alike. The increasing reliance on meta-analyses in all fields of psychology introduces a new frontier in understanding how the role of researcher belief, degrees of freedom and transparency in decision-making processes shape study outcomes. This new frontier requires critical self-reflection and a commitment to adopting practices which reduce the influence of bias, wherever possible.

Chapter 5

Experimental psychology's blind spot: Software validation

> Human-operant research with computer-based tasks has included little or no description of rigorous validation procedures for the experimental apparatus (i.e., the software used in the experiment). This omission, combined with a general lack of guidance regarding how to thoroughly validate experimental software, introduces the possibility that ... researchers may insufficiently validate their computer-based apparatus
>
> (Smith & Greer, 2022, p. 389)

5.1 Chapter Overview

This chapter focuses on one of the concerns raised in Chapter 2 about experimental psychology: software validation. Psychology experiments often use software when testing their participants, such as OpenSesame, PsyToolkit, E-Prime, and online computer-based research via crowd-sourcing websites, such as Amazon's Mechanical Turk. As highlighted in Chapter 2, mitigating experimenter fraud is essential, and software validation serves is a vital procedure of any psychology experiment to prevent manipulation by rogue experimenters or participants and to guard against error. Ensuring software validation is critical in psychology experiments, not only to confirm the validity of the data collected, but also to guard against any potential for manipulation or fraudulent behaviour, subtle or otherwise.

Given the contentious nature of parapsychological research and the inherent risks of fraud, software validation is a critical aspect of any parapsychology experiment. Yet, most psi ganzfeld studies use self-created software, independent to other researchers and are susceptible to critique about insufficient randomisation and bias checks. This chapter details the software validation conducted for KPU Study 1074, after data collection was completed, to assess whether the study's significantly positive results could have resulted from subtle flaws in the software that unintentionally contributed to the study's success.

5.2 Experimental psychology's blind spot

As mentioned in chapter 2, software validation is a crucial yet often overlooked aspect of psychology experiments. It tends to receive minimal attention in final reports, usually only briefly mentioned in the methods section. There are two primary concerns regarding experimental software that have not been adequately addressed in the literature: the standardisation of software and the validation of software to prevent manipulation and detect any artefacts in the software.

The first concern involves the lack of standardised experimental software. Behavioural and experimental paradigms suffer from a lack of coordinated efforts to develop and implement standardised software. This lack of coordination can result in divergent implementations of conceptually identical tasks, leading to error-prone code and making replication difficult (Sochat et al., 2016). Since reproducible research depends on using equivalent experiments, some initiatives, like the "Experiment Factory," have been developed to create a large, accessible, and open-source repository of standardised experiments (Sochat et al., 2016). While this is an important issue, it falls outside the scope of this thesis.

The second concern, and the focus of this chapter, relates to the validation of experimental software used by participants to prevent manipulation and fraud, and detect any errors in the software. Software validation is usually one of the final steps in the experimental design process, ensuring that the software functions as intended (S. W. Smith & Greer, 2022). This is often referred to as "black box testing," where the final version of the software is tested to ensure it produces appropriate outputs (such as accurate data collection) based on the inputs (such as user responses) (Smith & Greer, 2022). Standard validation typically involves testing the basic functionality of the software and removing any "bugs" before assessing other components dependent on these functions.

While software developers, in general, primarily focus on functionality and performance, they often overlook security concerns, which are not typically part of their decision-making process (Oliveira et al., 2014). Developers usually assume common cases for the inputs their code will receive and the potential states the program might reach, but they often do not consider uncommon cases that could be exploited by a skilled adversary (Oliveira et al., 2014). Despite this, these concerns have received little attention in the experimental psychology literature (Kennedy, 2014b, 2016; S. W. Smith & Greer, 2022). While inconsistent experimental results in psychology are often attributed to differences in

experimental procedures or subject populations, these inconsistencies are rarely considered to be the result of fraud (Kennedy, 2014b, 2016). As discussed in Chapter 2, the absence of thorough or standardised software validation can lead to artefacts arising from programming errors or oversights, which can compromise the validity of the research (Kennedy, 2016; Smith & Greer, 2022).

In parapsychology, a notable example involves the discovery of a software artefact after the completion of a psi study, as discussed by Watt and Brady (2002). Despite initially performing manual checks and randomisation checks on their software, the researchers observed an unexpectedly large psi effect in their predicted direction. As the effect was unexpectedly high, the authors conducted more systematic checks and subsequently identified a software artefact. The authors emphasised the potential influence of experimenter expectancy effects, suggesting that they might not have detected this artefact if the results had more closely aligned with their expected outcomes (Watt & Brady, 2002). This highlights the importance of software validation as a critical step in ensuring the validity and replicability of experimental findings. Given the contentious nature of the phenomena studied in parapsychology, rigorous software validation procedures are particularly important to prevent potential manipulation or fraud (Kennedy, 2014b, 2016).

This chapter reports on the software validation of the recently completed psi precognition ganzfeld study, Koestler Parapsychology Unit Study 1074. The study found evidence supporting the psi hypothesis, with a hit rate of 30%, which was significantly different from the mean chance expectation of 25% (exact binomial $z = 1.71$, $p = .043$, one-tailed). The objective of the software validation was to ensure that the randomness procedures incorporated in the study software were genuinely random and free from any subtle biases that could arise from participant usage. Such biases might have caused the software to act non-randomly, potentially inflating the hit scores or showing a preference for certain target pools or clips, which could have contributed to the study's significantly positive results.

5.2.1 Software validation example: KPU Study 1074

5.2.2 Existing security measures

Koestler Parapsychology Unit Study 1074 incorporated a variety of methodological features to reduce the opportunity for experimenter and participant fraud, as detailed in the

study preregistration document https://edin.ac/4eYch2R. As stated by Kennedy (2014a), he argues there are two ways in which fraudulent researchers can manipulate a study, data manipulation and analysis manipulation chapter 2. This chapter focuses on the former, as chapter 6 addresses the latter form of manipulation. KPU study 1074 included measures to prevent data manipulation, as detailed below.

Data manipulation

Data manipulation is overt fraud by an experimenter. While this type of misconduct is a recognised threat to the integrity of research findings, it often remains difficult to detect due to the interpersonal costs and embarrassment involved in accusing colleagues of such behaviour (Kennedy, 2014b; Stroebe et al., 2012). There are documented cases of researcher fraud within parapsychology (Braude, 2021; Palmer, 2016; Rhine, 1975). To minimise the opportunity for experimenter fraud and to safeguard the integrity of the collected data, KPU study 1074 incorporated several methodological features:

1. Restricted access to the ganzfeld lab: The lab was in a secure section of a University of Edinburgh building with electronic ID card access requiring prior authorisation. Access to the study room was controlled by keypad code entry.

2. Multiple experimenters: Fifteen experimenters conducted the data collection, with each experimenter running one session per participant, for up to 20 trials per experimenter. Experimenters worked in blocks, minimising the risk of collusion or rogue behaviour.

3. Local and remote storage of experimental data files: Local computer data were automatically date and time stamped. The same data files were sent to a remote University of Edinburgh server, which was also time stamped upon receipt, preventing alteration by any experimenter. The server recorded the IP address of incoming data, detecting data sent from any source other than the designated experiment PC. This remote server was managed by an IT staff member unaffiliated with the KPU or the research project.

4. Multiple session data outputs: The experimental software saved data at three time points during each session, both locally and remotely. The first data capture occurred at the start of the session, recording details such as the session type (informal or formal),

experimenter ID, participant ID, and demographic information (date and time). The second occurred when the participant logged their final target ratings, before the target was randomly selected. The third capture happened after the target was selected but before the experimenter or participant could knew the target's identity. All session data, both informal and formal, were logged in this manner.

5. True random number generator: A commercially available random number generator (TrueRNG3) was used to ensure unbiased target selection. This RNG was tested at the start of each experimental wave to verify the lack of bias in the relative frequency of target pool selections (1/50), target clip presentations (1, 2, 3, 4), and target selections (1/200).
6. Audio recording of mentations: Each mentation was audio-recorded for transcription and session verification. The experimenter announced their name, session ID, date, and time at the start of each recording to match it with other session records.

7. Manual logbook: Experimenters maintained a manual logbook to crosscheck session data. The logbook was glue-bound and hand-numbered by the principal investigator, making any missing or torn pages evident. Experimenters manually logged whether the session was informal or formal, session details (date and time), experimenter name, participant ID, order of target presentation, submitted ratings, and session results (hit or miss and target identity). The manual log served as the "true responses" record, cross-referenced with the "recorded responses" in the digital data files (Smith & Greer, 2022).

8. Independent verification: After data collection was completed, an independent programmer from the School verified the locally stored data against the data sent to the server to ensure no differences or manipulations.

Before formal testing began, 24 informal ganzfeld sessions were conducted to check for software "bugs." Six hits were obtained in the informal sessions, consistent with the MCE of 25%. All data from these sessions were logged both locally and remotely and recorded in the manual logbook. Additionally, before conducting any formal sessions, each experimenter completed at least one informal session to ensure they were comfortable with the software and procedure. This informal session tested the software in the same experimental context in

which it would be used by the experimenter and participant, as recommended by Smith and Greer (2022).

The next section details the additional software validation conducted after data collection. Although this was not part of the original study design (a limitation discussed later), it serves as an example of further validation beyond basic user and acceptance testing.

5.4 Method

The additional software validation was conducted by an independent individual affiliated with the University of Edinburgh but without any connection to the department, the research team, or involvement in the experimental software's design or coding. The code for the software validation process is publicly available at https://github.com/drpawelo/study_testing_precognition. Statistical analyses were conducted in JASP (JASP Team, 2024).

My role in the software validation process consisted of explaining the study methodology to the programmer, detailing the roles of both the experimenter and the participant, and clarifying the inputs they would provide to the experimental software. I also outlined the aim of the validation process to ensure the programmer fully understood the objectives and procedures. The goal was to simulate participant inputs ("puppeteering") on the experimental software under conditions identical to those of the original study, including the use of the same TrueRNG3 random number generator. The validation aimed to assess four features which involved randomisation:

1. Random Selection of Target Videos: To determine whether the software's random selection mechanism (TrueRNG3) was functioning correctly over a larger sample size than the randomisation checks conducted throughout the experiment. This involved verifying that each video in the randomly selected pool had an equal 25% chance of being chosen as the target.

2. Alignment of Video Ratings with Random Selection: To check if, when video ratings were distributed randomly, the frequency of matching the highest-rated video to the randomly selected target by TrueRNG3 aligned with the expected probability of 25%.

3. Random Selection of Video Pools: To test whether each of the 50 possible video pools had an equal chance of being selected (1/50), ensuring there was no bias in the randomisation process.

4. No Bias in Highest-Rated Video Clips: To verify that there was no bias in which of the four video clips was the highest-rated video clip when ratings were assigned randomly.

The software validation involved creating a remote-controlled "participant" to run many simulated sessions. These simulated participant inputs were then compared against the expected chance results to identify any potential biases within the experimental software. The validation script utilised Python's *pyautogui* library https://github.com/asweigart/pyautogui, which simulated participant actions by controlling the mouse and keyboard as needed. The library also generated random guesses using Python's *random.shuffle()* function when the participant was expected to make decisions. The script was repeatedly opened, closed, and restarted, running continuously over four weeks without any experimenter supervision.

Two series of software validation sessions were conducted: "long" and "short" sessions. The long sessions, designed to replicate real experimental conditions, ran at real-time speed and had an average duration of 58 minutes, completing 280 trials in total, to match the 'real' number of studies. The short sessions, which reduced the duration of the video and audio clips to one second each while keeping the rest of the study unchanged, averaged 2 minutes per session and performed 5,549 trials.

5.5 Results

The validation revealed no bugs or discrepancies, confirming that the experiment software performed as expected under both human and script-controlled conditions. The script-based validation showed that the software's randomisation and selection processes operated correctly even without human observation, enhancing the credibility of the study outcomes. This thorough validation process reinforced our confidence in the software's reliability and the integrity of the collected data. Results split by session type are detailed below.

5.5.1 Long sessions

Random selection of target video

The simulation sessions confirmed that each video in the randomly selected target pool had an equal chance of being chosen as the target clip. Specifically, video clip number 1 was selected 66 times, the second clip 77 times, the third clip 55 times, and the fourth clip 71 times. A multinomial test, with the null hypothesis that each clip had an equal probability of 0.25, failed to reject the null hypothesis ($\chi^2(3) = 1.17, p = .76$; see Table 5.1). The descriptive plot, including 95% confidence intervals, is presented in Figure 4.1a.

Table 5.1

Observed Proportion of Video Selected as Target Clip

Video Number	Observed	Expected: H0 (a)	95% Confidence Interval	
			LL	UL
1	0.24	0.25	0.19	0.29
2	0.28	0.25	0.22	0.33
3	0.24	0.25	0.19	0.29
4	0.25	0.25	0.24	0.31

Note. Confidence intervals are based on independent binomial distributions.

Number of hits and misses

For the second feature, the simulation resulted in 59 hits out of 280 sessions. With an expected hit probability of 25%, the software validation confirmed that the hit rate was not significantly different from this expectation when target ratings were randomly allocated ($\chi^2(1) = 2.31, p = .13$). The simulation yielded a hit rate of 21.1% (95% CI [16.4, 26.3]), while 78.9% of the simulated sessions were misses (95% CI [73.7, 83.6]). The descriptive plot, including 95% confidence intervals, is shown in Figure 4.1b.

Random selection of video pool

For the third feature, the simulation assessed whether each video pool had an equal chance of being selected. The results indicated that the video pools were selected equally ($\chi 2(48) = 30.80, p = .98$).

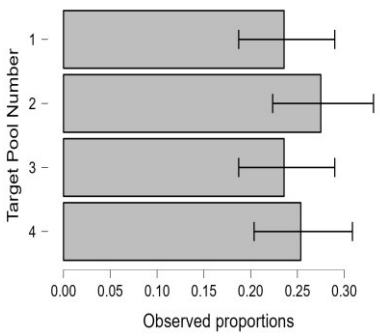
Highest rated video number

For the fourth feature, the aim was to verify that each video clip had an equal chance of being rated the highest when ratings were allocated randomly. Video number 1 was rated the highest 73 times, the second clip 83 times, the third clip 56 times, and the fourth clip 68 times. A multinomial test failed to reject the null hypothesis that each video clip had an equal probability of being rated the highest ($\chi^2(3) = 5.40, p = .15$). The descriptive plot, including 95% confidence intervals, is shown in Figure 5.1c.

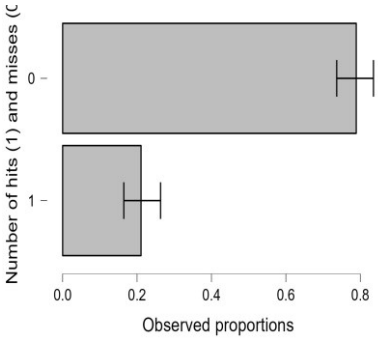
Figure 5.1

Observed Proportions for Simulated Long Sessions with 95% Confidence Intervals

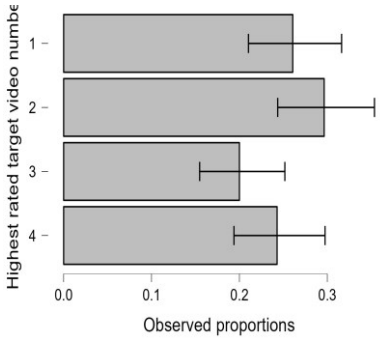
(a) Proportion each Video Clip was Target



(b) Proportion of Hits and Misses



(c) Proportion each Video Clip was Highest Rated



5.5.2 Short sessions

Random selection of target video

The simulation assessed whether each video number in the pool had an equal chance of being selected as the target. The first video was selected 1,399 times, the second video 1,394 times, the third video 1,385 times, and the fourth video 1,371 times. A multinomial test failed to reject the null hypothesis that each video clip had an equal probability of being

selected as the target ($\chi^2(3) = 0.33$, $p = .96$). The observed proportions and 95% confidence intervals are shown in Figure 4.2a.

Number of hits and misses

The puppeteering software produced a total of 1,364 hits out of 5,549 sessions, resulting in a hit rate of 24.6% (95% CI [23.5, 25.7]). This hit rate was not significantly different from the expected probability of 25% ($\chi^2(1) = 0.52$, $p = .47$). The observed proportions and 95% confidence intervals are shown in Figure 4.2b.

Random selection of video pool

the software validation assessed if the video pools had equal chance of being selected. A multinomial test was conducted and failed to reject the null hypothesis that the video pools had an equal chance of being selected ($\chi^2(49) = 28.68$, $p = .99$).

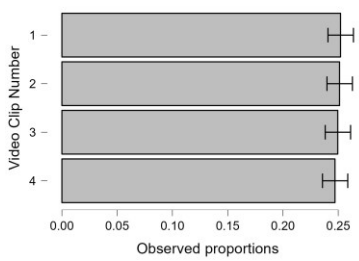
Highest rated video number

The puppeteering software rated the first clip as the highest 1,351 times, the second clip 1,447 times, the third clip 1,347 times, and the fourth clip 1,404 times. A multinomial test found that the video clips within each pool were rated equally as the highest by the simulated participants ($\chi^2(3) = 4.89$,$p = .18$). The observed proportions and 95% confidence intervals are shown in Figure 5.2.

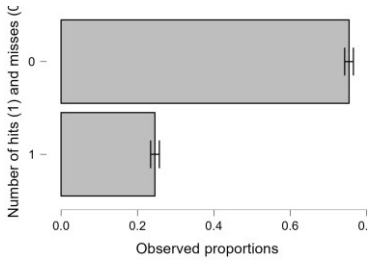
Figure 5.2

Observed Proportions for Simulated Short Sessions with 95% Confidence Intervals

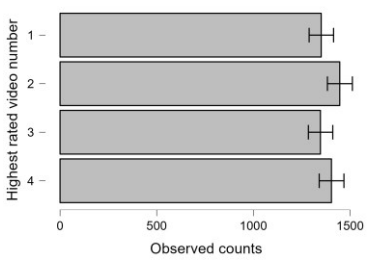
(a) Proportion each Video Clip was Target



(b) Proportion of Hits and Misses



(c) Proportion each Video Clip was Highest Rated



5.6 Discussion

This chapter reports the first (known) extensive software validation of psi ganzfeld experimental software. This validation procedure reduces concerns that the positive results were influenced by subtle programming biases or a malfunctioning random number generator. Alongside other methodological safeguards—such as data duplication, manual record-keeping, and various security features—this independent software validation offers additional assurance that the experimental software did not introduce bias in target generation.

In general, experimental psychologists should validate their software and standardise their experiments, as replicability fundamentally depends on these practices (Oliveira et al., 2014). Even while focusing on psi research, Kennedy (Kennedy, 2014a) emphasised that "experimenter fraud should not be easy and tempting..." (p. 8) and recommended involving multiple experimenters in a study, storing duplicate copies of randomisation and session data, and having experimenters check and observe one another. These recommendations are straightforward and could be extended to other areas of psychology as basic principles. In fact, such practices are already standard in pharmaceutical research (Kennedy, 2014b, 2017).

5.6.1 Limitations

Despite these extensive validation efforts, several limitations must be acknowledged. First, the timing mechanisms of the software were not validated. This includes verifying whether audio files (such as relaxation and white noise periods) and target clips (each approximately one minute long) adhered to their intended duration and whether there were any subtle discrepancies in their start times or lengths. Discrepancies could potentially introduce bias if certain pools or targets were presented at different times across trials. As suggested by Smith and Greer (2022), a frame-by-frame analysis at multiple points can ensure synchronisation throughout the experiment. Unfortunately, such an analysis was not performed, and any inadvertent changes in the lengths of the audio files or videos may have affected the software's behaviour and introduced unintended bias.

Second, the research team had access to the source code of the experimental software. While implementing kiosk mode—which restricts a device to running a single application or a limited set of functionalities to prevent unauthorised access to the underlying operating

system—was considered to enhance security, it was not feasible due to limitations of the software platform (OpenSesame). However, the use of multiple data storage methods (local, remote, and manual logbooks) helped ensure data integrity and mitigated concerns about potential manipulation or bias. Further, having multiple experimenters working for set time periods strengthened security and allowed for easier identification of any suspicious activity.

Finally, it is important to note that the software validation was conducted after data collection had been completed, relying on "puppeteered" sessions that used simulated participant inputs to make decisions and actions. While this approach successfully validated the random number generator and overall functioning of the experimental software, future researchers should conduct such checks before formal data collection begins. Early validation would allow for the identification and correction of potential issues or biases before the experiment's commencement, providing greater assurance of study integrity. Moreover, if a software artefact had been present, early detection would have allowed for its correction, avoiding unnecessary expenditure of time, money, and effort on a study that could otherwise be invalidated by a software error.

As exemplified, Watt and Brady (2002) did not discover a software artefact in their experimental setup until after their first study had concluded. During the pilot sessions, experimenters, and other researchers, who were familiar with the study design, interacted with the software in ways that did not reflect the actual behaviours and inputs of the study participants. As noted by Oliveira et al. (2014), the experimenters only inputted common cases and scenarios they expected, rather than accounting for the diverse range of inputs that participants could potentially provide. This approach left the artefact undetected until after the study was completed.

Future software validation efforts for psi ganzfeld and similar psychological experiments should consider simulating validation sessions that incorporate movements and behaviours generated by human participants to enhance the ecological validity of the findings. The software validation in this study relied on random guesses for decision-making, which does not accurately reflect human behaviour, as humans typically do not act randomly. Introducing randomness into the validation process might obscure the detection of non-random patterns, which is often the focus of the validation itself. To improve future validations, researchers may adopt the recommendations of Smith and Greer (2022) to recruit participants in smaller increments and evaluate the data iteratively. During these pilot sessions, researchers could

record participant behaviours—such as mouse movements (Freeman & Ambady, 2010), excessive clicking, result changes, or repeated viewing of target clips. Randomly sampling these recorded actions for use in validation simulations would provide a more realistic test of the software.

In the field of parapsychology, experimental researchers are primarily concerned with the potential for fraud and the need for stringent security measures, given the controversial nature of the phenomena under investigation (see Dalton et al., 1996; Kennedy, 2017; Rhine, 1975). Verifying the software used in these studies is crucial to ensure that the positive results of KPU Study 1074 were not due to software artefacts, manipulation, or fraudulent behaviour by the researchers. Furthermore, the enhanced security measures implemented have not led to non-significant outcomes in psi ganzfeld studies (e.g., Milton & Wiseman, 1999). However, as discussed in chapter 4 and chapter 6, the lack of a standardised psi ganzfeld framework and experimental software raises concerns about the comparability of different ganzfeld studies.

5.7 Conclusion

This chapter presents the first known software validation of a psi ganzfeld experiment, which aimed at ensuring that the results were not influenced by subtle biases or non-randomness in the experimental software. The software validation, along with other security measures, did not impose significant logistical or time burdens on the study. Therefore, such validation procedures should be incorporated into psychological experiments to minimise the potential for bias, undetected errors and reduce the risk of experimenter fraud. Due to the contentiousness of psi experiments, insufficient randomisation and poor software security are common concerns (Dalton et al., 1996; Wiseman et al., 1994). But the psi ganzfeld can be an example of continuous methodological improvement in search of approval by the wider psychology community by using scientific methods to avoid accusations of questionable research practices.

Chapter 6

Specification curve analysis of psi ganzfeld data

6.1 Chapter overview

This chapter uses data from the Koestler Parapsychology Unit (KPU) Study 1074 to perform a specification curve analysis (SCA). As discussed in Chapter 2, SCA, like other multiverse-type analyses, demonstrates the numerous ways researchers can structure their statistical approaches and prevent analysis manipulation. Using data from a recently completed precognition study, this chapter examines numerous ways this single dataset can be constructed and analysed. The chapter begins with a brief overview of the data collected in KPU Study 1074, detailing the rationale for selecting participants and measuring specific traits, hypotheses and planned analyses. These decisions were guided by previous literature and recommendations from other psi ganzfeld studies, which have demonstrated diverse outcomes and analytical methods. This diversity highlights the lack of a singular, definitive approach to modelling and analysing such data leading to questions of selective reporting and other QRPs.

The chapter emphasises the thesis topic that despite the use of preregistration and the application of optimised practices based on prior guidelines, a wide range of potential approaches to constructing and analysing the data remains. Even when constrained to a limited set of variables from a single dataset, the specification curve analysis demonstrates the myriad ways in which researchers can construct their statistical models.

6.2 Building on lessons learned: KPU Study 1074

This chapter utilises data collected from the Koestler Parapsychology Unit (KPU) Study 1074, which was developed in line with the latest guidance and recommendations for conducting a psi ganzfeld study. Study 1074 developed the paradigm and methods used in an earlier exploratory study using a precognition ganzfeld paradigm (Watt et al., 2020), which produced a result significantly supporting the psi hypothesis. The primary objective of Study

1074 was to test a confirmatory hypothesis that the precognition ganzfeld is conducive to psi effects.

A summary of KPU Study 1074 is provided in this chapter, including key aspects such as the selection criteria for participants, the measurements and questionnaires utilised, and the data collected. This establishes a baseline against which the specification curve analysis can be compared, illustrating the multitude of ways in which the study data can be collated, structured, and analysed. For a comprehensive description of the study procedure and additional details, refer to the pre-registration document https://edin.ac/4eYch2R and the final report for the funding organisation https://edin.ac/4e8Wo8t.

6.2.1 Background

KPU Study 1074 was designed according to the latest guidance for psi ganzfeld research, incorporating several methodological features to enhance rigour and validity. The study used a large sample size of 240 participants, which is substantial for ganzfeld studies[9]. This large sample of selected participants provided the study a power of 98% to detect a hit rate of 36.5%. Participants were selected, with a focus on recruiting creative individuals, as prior research has indicated that creative participants tend to achieve higher hit rates in psi experiments (Baptista et al., 2015; Dalton, 1997; Schlitz & Honorton, 1992). Further, the experimenters were selected to have self-reported neutral or positive expectation towards the study obtaining the psi hypothesis.

The study was preregistered, and the study design also made the study eligible for inclusion in a prospective meta-analysis (Watt & Kennedy, 2017) and included numerous methodological features to reduce fraud and bias. The study employed a precognitive design, where the target clip was selected *after* the participant had rated all four possible target clips. The use of 15 experimenters minimised any subtle biases or influences that might arise from a single experimenter and prevented tiring out experimenters. A true random number generator (TrueRNG3) was used to randomise the selection of the target pool, the order of presentation, and the final target selection, ensuring all processes were genuinely random. For data security, information was recorded in multiple formats, including physical documents, local computer files, and files stored on a remote server inaccessible to the experiment team.

[9] Fifty studies which had the number of participants extracted in Tressoldi (2019) had an average sample size of 35.

Cross-checking confirmed that no data manipulation occurred. Additionally, the study included software validation procedures (detailed in chapter 5) to confirm that all computational and randomisation processes were performing without bias.

6.2.2 Participants and assessments

Participants for the study were recruited using opportunity and snowball sampling methods. They were screened for eligibility based on their responses to a demographic questionnaire (Participant Information Questionnaire; PIQ). To qualify for inclusion, participants had to answer "Yes" to the question, "Do you engage in any artistic/creative activities?" and rate their creative/artistic ability as 3 or higher on a 5-point scale, where 1 indicated "low" and 5 indicated "high." Additionally, participants needed to report at least one of the following: a practice of a mental discipline on a weekly basis or more frequently, prior psi experiences, or a belief in psi (such as ESP or psychokinesis). Participants were excluded if they were under 18 years of age, were not within travelling distance of Edinburgh, or had a current mental health disorder. Eligible participants then completed a series of validated questionnaires. These included:

Australian Sheep-Goat Scale (ASGS), which assesses paranormal beliefs, psychic abilities, and experiences. Participants rated each statement on the ASGS as either 0 ("false"), 1 ("uncertain"), or 2 ("true"), with total scores ranging from 0 to 36 (Thalbourne, 2010).

Inventory of Creative Activities and Achievements (ICAA), which measures creative activities and achievements across eight domains: literature, music, creative cooking, visual arts, arts and crafts, performing arts, sports, and science and engineering. The ICAA has two sub-scales: the Creative Activities subscale (CAct) and the Creative Achievements subscale (CAch) (Diedrich et al., 2018). The creative activity consists of 6 items per domain (e.g., "Wrote a short literary work" for the literature domain), rated on a 5-point Likert-type scale (0 = "Never," 1 = "1-2 times," 2 = "3-5 times," 3 = "6-10 times," 4 = "More than 10 times"). Scores are averaged across the 6 items to yield a domain-specific score and summed across all eight domains to generate a domain-general score. The creative achievement subscale measures 11 levels of creative achievement, scoring from 0 to 11 per level. These scores are summed to yield a domain-specific score ranging from 0 to 55, and further summed across all domains for a general score.

Runco Ideational Behavior Scale (RIBS), which assesses creative ideation by asking participants to rate 23 items (e.g., "I have ideas about new inventions or about how to improve things," "I am able to think about things intensely for many hours") on a 5-point scale (1 = "never" to 5 = "very often"). The overall RIBS score is computed by averaging the ratings across all 23 items, resulting in a value between 1 and 5 (Runco et al., 2001). Once these additional questionnaires were complete, participants were invited to the ganzfeld lab to conduct one session.

Participant demographics and questionnaire descriptive statistics are shown in Table 6.1. Out of the 240 participants, 183 were female, 51 were male, 4 identified as non-binary/third gender, one participant preferred not to say and one gender was missing. As explained in chapter 7, not all participants produced a mentation and some spoke in languages other than English and were not transcribed, hence the lower numbers for the language variables. The language variables were constructed using LIWC2015 software, as detailed in chapter 7.

Table 6.1

KPU Study 1074 Participant Demographics and Questionnaire Descriptive Statistics

	Valid	Missing	Mean	SD	Min	Max
Age	239	1	28.80	13.46	18.00	84.00
ASGS	239	1	18.64	7.97	0.00	36.00
Creative Activity (General)	239	1	10.73	4.71	1.50	26.17
Creative Achievement (General)	239	1	83.52	50.71	4.00	270.00
RIBS	239	1	3.62	0.61	1.78	4.83
Word Count	177	63	513.89	511.35	2.00	2942.00
Analytic	177	63	68.00	29.09	4.31	99.00
Clout	177	63	33.99	19.82	1.00	92.33
Authentic	177	63	68.26	28.92	1.00	99.00

6.2.3 Hypotheses and planned analyses

The study had five hypotheses. The first hypothesis was confirmatory and all subsequent were exploratory, with all examined at the 5% significance level. The dependent variable for the exploratory hypotheses is the session *z*-score which is defined as:

$$z = \frac{x_i - \bar{x}}{\sqrt{s_x / n}}$$

where x_i is the target rating, $\bar{x}$ is the mean of all target ratings, s_x is the standard deviation of all target ratings and n is the number of ratings (4).

H1: Participants will correctly identify the randomly selected target clip at greater than mean chance expectation (25%). Tested using exact binomial probability ("k out of n") of the number of hits out of 240 trials where the probability p is the probability that the outcome will occur on any particular occasion (0.25).

H2: Paranormal belief (measured by ASGS) will correlate positively with ganzfeld task performance (session *z*-score).

H3a: Creativity (measured by ICAA) will correlate positively with ganzfeld task performance (session *z*-score).

H3b: Creativity (measured by RIBS) will correlate positively with ganzfeld task performance (session *z*-score).

H4: Quantity of ganzfeld mentation (number of words spoken by participant) will correlated with ganzfeld task performance.

H5: Content of ganzfeld mentation (measured using LIWC2015) will correlate with ganzfeld task performance.

6.2.4 Results

The results presented here for hypotheses 2 and 3 differ in sample size due to additional data checks and a quantitative language analysis conducted after the funding report submission. However, these results still use the same planned analyses (correlations) as defined in the study preregistration. The final funding report did not include results for hypotheses 4 and 5 because the mentation analysis had not yet been completed at that time.

The current results adhere to the planned analyses, following the study preregistration document and the predetermined *a priori* modelling choices.

As reported in the final funding report, the confirmatory hypothesis (H1) was supported, as the study produced 72 hits out of 240 sessions (30% hit rate, $p = .043$, one-tailed, $z = 1.714$). Adhering to the preregistered planned analyses, the onetailed correlations assessing hypotheses 2, 3a and 3b are shown in Table 6.2. None of the variables had a significant relationship with session outcome. Table 6.3 shows the two-tailed Spearman's correlations with mentation word count and four summary variables, discussed in chapter 7. Hypotheses 4 and 5 are also not supported as there are no significant correlations with session outcome.

6.3 Specification curve analysis (SCA)

KPU Study 1074 was preregistered and designed according to recommendations from previous research (namely Baptista et al., 2015; Kennedy & Watt, 2018; Storm et al., 2010b) to create a methodological sound (or as close) precognition psi ganzfeld study. However, even with preregistration, hypothesis tests are often not specified in a way that eliminates flexibility in statistical analysis (Scheel et al., 2021). The preplanned analyses defined in study 1074 can be constructed in a variety of different ways. For example, for the exploratory hypotheses looking at mentation content, the

Table 6.2

Spearman's Correlations of Creativity Measures, Paranormal Belief and Session Outcome

Variable		Session z-score	Creative Ideation	Creative Activity (General)	Creative Achievement (General)	Paranormal Belief
Session z-score	Spearman's rho	—				
	p-value	—				
	Upper 95% CI	—				
	Lower 95% CI	—				
Creative Ideation	Spearman's rho	-0.052	—			
	p-value	0.788	—			
	Upper 95% CI	0.071	—			
	Lower 95% CI	-0.180	—			
Creative Activity (General)	Spearman's rho	-0.019	0.438***	—		
	p-value	0.613	<.001	—		
	Upper 95% CI	0.108	0.541	—		
	Lower 95% CI	-0.133	0.310	—		

Creative Achievement (General)	Spearman's rho	-0.021	0.338***	0.647***	—	
	p-value	0.625	<.001	<.001	—	
	Upper 95% CI	0.103	0.452	0.713	—	
	Lower 95% CI	-0.140	0.209	0.562	—	
Paranormal Belief	Spearman's rho	-0.023	0.282***	0.251***	0.170**	—
	p-value	0.639	<.001	<.001	0.004	—
	Upper 95% CI	0.100	0.406	0.369	0.297	—
	Lower 95% CI	-0.145	0.156	0.128	0.040	—

Note. All tests one-tailed, for positive correlation. Confidence intervals based on 1000 bootstrap replicates. * p <.05, ** p <.01, *** p <.001, one-tailed

Table 5.3

Spearman's Correlations of Mentation Content and Session Outcome

Variable		Session z_score	Word Count	Analytic	Clout	Authentic	Tone
Session z-score	Spearman's rho	—					
	p-value	—					
	Upper 95% CI	—					
	Lower 95% CI	—					
Word Count	Spearman's rho	0.029	—				
	p-value	0.705	—				
	Upper 95% CI	0.174	—				
	Lower 95% CI	-0.119	—				
Analytic	Spearman's rho	-0.115	-0.350***	—			
	p-value	0.127	<.001	—			

	Upper 95% CI	0.043	-0.214	—			
	Lower 95% CI	-0.261	-0.481	—			
Clout	Spearman's rho	-0.011	-0.188*	0.602***	—		
	p-value	0.880	0.012	<.001	—		
	Upper 95% CI	0.149	-0.035	0.690	—		
	Lower 95% CI	-0.160	-0.327	0.477	—		
Authentic	Spearman's rho	0.064	0.198**	-0.495***	-0.732***	—	
	p-value	0.400	0.008	<.001	<.001	—	
	Upper 95% CI	0.208	0.351	-0.368	-0.643	—	
	Lower 95% CI	-0.085	0.039	-0.604	-0.798	—	
Tone	Spearman's rho	0.017	0.199**	-0.166*	-0.053	0.127	—
	p-value	0.825	0.008	0.027	0.483	0.092	—
	Upper 95% CI	0.167	0.361	-0.013	0.104	0.286	—
	Lower 95% CI	-0.131	0.039	-0.317	-0.214	-0.025	—

Note. Confidence intervals based on 1000 bootstrap replicates. * p <.05, ** p <.01, *** p <.001

6.4. Method

LIWC2015 software provides over 90 variables -- a researcher could use any one, or combination, of these variables to assess this single hypothesis. Researchers frequently face what Gelman and Loken (2016) describe as a "garden of forking paths", where analytical decisions can be arbitrary, even in preregistered studies (Simonsohn et al., 2020). This often leads to the reporting of a single analysis or a limited set of variations, which may not fully capture the entire range of plausible analytical approaches (Del Giudice & Gangestad, 2021; Plessen et al., 2022; Voracek et al., 2019).

Additionally, when different researchers analyse the same data with the same hypothesis, they construct different models at significantly different conclusions (Breznau et al., 2022). To address this flexibility and subjectivity, there has been a growing adoption of multiverse-style analyses (as discussed in chapter 4), which estimate effects across a range of plausible specifications to identify hidden degrees of freedom and reduce false positives (Del Giudice & Gangestad, 2021; Steegen et al., 2016). One approach is the specification curve analysis (Simonsohn et al., 2020). Specification curve analysis involves defining a list of reasonable model specifications, analysing all these specifications and visually presenting the results to identify crucial decisions in the specification process. The inferential test collectively assess whether the model specifications reject the null hypothesis that the effect of interest does not exist (Simonsohn et al., 2020). This approach provides a more comprehensive understanding of the robustness of findings by considering a broader range of analytical choices available to researchers.

Specification curve analysis is applied below, using the data from KPU 1074. The application of the SCA in this context is intended primarily for exploratory and illustrative purposes, rather than for effect estimation. Accordingly, the SCA presented below is limited in terms of possible specifications to highlight that, even with such restrictions, the number of plausible combinations remains extensive.

All analyses were conducted in *RStudio* 2024.04.2 Build 764, using the *specr* package (Masur & Scharkow, 2020). The R code can be found at https://github.com/yeloopa/PhD. The main

focus of this SCA is examining exploratory hypotheses 2 to 5, as they are not defined in a way which eliminates flexibility in the variable coding, construction and statistical analysis.

6.4.1 Identifying specifications

The variables, domains, operationalisation, and construction of variables used in this analysis are summarised in Table 6.4. Cells which contain an asterisk denote variables in their raw format, as collected in Study 1074. All other constructions were determined and constructed by myself. To further illustrate the flexibility in conducting such analyses, the last column presents alternative ways the variables could have been constructed. Other relevant decisions, such as variable transformations, interpretability, or handling of missing data, were not considered, as the primary focus of this SCA is on the decision-making processes related to participant trait variables and alternatives, which is demonstrated in the psi ganzfeld and broader psychological literature (Honorton, 1997; Rauvola & Rudolph, 2023; Steegen et al., 2016).

The exploratory hypotheses (2-5) in KPU study 1074 focused on the relationships between paranormal belief, creativity, and mentation content and session outcome. Two dependent variables were analysed: the session z-scores, which are commonly used as a measure of session outcome and binary hit rates, which has been used as an outcome measure in previous reports (Milton, 1997a).

Eight independent variables were included in the specification curve analysis. The mentation variables, word count, and four summary variables from the LIWC2015 (Analytic, Tone, Clout, Authenticity), were incorporated in their original forms. These four LIWC2015 summary variables were selected to limit the analysis to a manageable subset, given the substantial number of potential LIWC2015 outputs. Additionally, the RIBS score and two general measures (Activity and Achievement) of the ICAA were included as predictors, also in their original forms, consistent with the analysis in the final funding report.

In the *specr* framework, including subsets is akin to interaction terms in a linear model, allowing for comparisons across different levels of factors and various combinations of predictors and controls. For this analysis, two subsets were created. The first is paranormal belief, using the ASGS score, which was recoded into a factor with two levels, "skeptical" and "believer". Participants who scored or below the mean were classified as "skeptical", while those score above the mean were classified as "believers". This method mirrors

recoding of trait questionnaires used by other researchers, where numerical variables are constructed into low and high categories, like Marcusson-Clavertz and Cardeña (2011) and Cardeña and Marcusson-Clavertz (2020). The second subset is created based upon the self-reported frequency of mental discipline practice from the PIQ. Originally an ordinal variable, it was recoded into a factor with levels, "regular" and "infrequent". This reflects the selection criteria of KPU 1074 where participant who practiced a mental discipline weekly or more were selected over those who practiced less frequently. Additionally, two control variables– self-reported gender and age (years)– were included in the analysis, as commonly controlled for in psychological research. Even when limiting the SCA to these variables and modelling choices, there is already 1,024 different ways of constructing a model. As shown in Figure 6.1 panel B, the decision tree becomes complex rather quickly with each node reflecting a decision point.

I conducted a second specification curve analysis, focusing exclusively on the session z-score as the outcome and using linear modelling for the model decision, while keeping all other variables constant. This approach was chosen because the majority of existing literature utilises the z-score to examine relationships between participant traits and characteristics.

Table 5.4

Study Variable Categories, Domains, Operationalisations and Variable Constructions in SCA

Variable	Domain	Operationalisation	Variable construction	Alternative valid constructions (not examined)
Dependent variable	Session outcome	z-score	*	
		hit rate (0/1)	Binary 0/1	Binary 0/1 where top two ratings combined and bottom two ratings combined. If target in the top two combined = 1, otherwise, 0.
Independent variable	Creativity	General Creative Achievement score (ICAA)	Numeric variable: Sum across all 8 creative domains*	Group by quartiles ("low" <.25 ≤ "mid" ≤ .75 <"high"); "High" ≥ mean/median, "Low" <mean/median
		General Creative Activity score (ICAA)	Numeric variable: Sum across all 8 creative domains*	As above
		RIBS	Numeric variable: mean value of responses to 23-item questionnaire*	As above
	Mental discipline	Self-reported "Have you ever practised any form of mental discipline...?" (PIQ)	Factor with two levels,* yes = 1, no = 0	
	Mentation features	Word count (LIWC2015)	Numeric variable: total number of words produced by participant*	Group by quartiles, ("low" <.25 ≤ "mid" ≤ .75 <"high"); "High" ≥ mean/median "Low" <mean/median
		Analytic (LIWC2015)	Numeric variable: Analytical thinking score 0-100*	As above
		Tone (LIWC2015)	Numeric variable: Emotional tone score 0-100*	As above
		Authentic (LIWC2015)	Numeric variable: Authentic score 0-100*	As above
		Clout (LIWC2015)	Numeric variable: Clout score 0-100*	As above
Statistical model	Linear regression Logistic regression			
Controls	Age	Self-reported (years)*	Numeric variable*	Group into different cohorts, 18-24 years, 25-34 years, 35-44 etc.
	Gender	Self-reported sex*	Factor with 4 levels: female, male, non-binary/third-gender, prefer not to say*	
Subsets	Mental discipline	Self-reported "How often do you practice [a mental discipline]?"	Factor with two levels: daily/several times a week/weekly = "regular", several times a month/monthly or less = "infrequent"	Keep as original Likert-type; split into three levels (regular, frequent, infrequent)
	Paranormal belief	ASGS	Factor with two levels: equal to and less than the mean = "skeptical", greater than the mean = "believer"	Keep as original numeric score; Group by quartiles, ("low" <.25 ≤ "mid" <= .75 <"high")

Note * indicates the raw, unchanged variable construction as collated in the study database.

6.5 Results

The SCA results suggest that that almost all model specifications produce a null result. Figure 6.2 displays the specification curve where each plotted point denotes an effect (i.e., regression coefficient) estimated by the model specification. These specifications are ranked by order of magnitude, with the colour denoting direction and significance. Red indicates a significant negative effect, grey indicates a non-significant effect and blue indicates a significant positive effect, all at the 5% significance level. Dashes on the curve plot (Panel A) are the 95% confidence intervals for the estimate of each model. The median effect size of all specifications is 0 (*min* = -3.12, *max* = 0.88) , and the sample sizes for specifications range from 21 to 239, with a median of 83. The specification curve and specification decision plots are presented individually in Appendix C to see in more detail.

Across the full set of model specifications, the variance is predominantly explained by residual error, which accounts for 62.60%. Next, is the choice of predictor, which explains 24.14% of the variance across different model specifications. This is followed by the decision surrounding the subsets, which contributes 8.69% to the variance. The type of model used and the choice of dependent variable account for 2.41% and 2.16% of the variance, respectively. The control variables (gender and age) accounted for 0% of variance across model specifications.

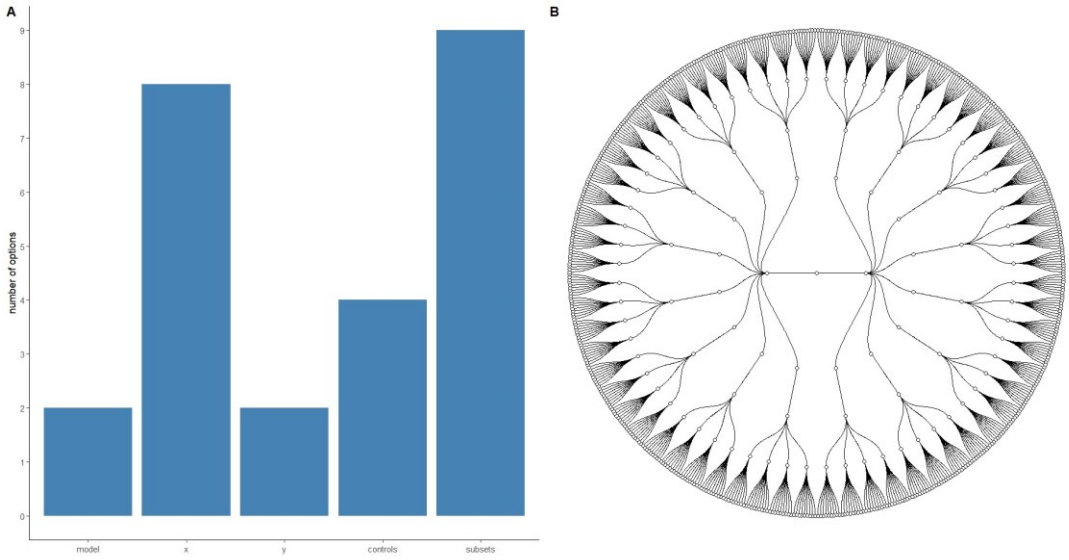
For the second SCA, the majority of model specifications produced null results, as shown in Figure 6.3. There was a total of 512 specifications, with median effect size of 0 (*min* = -3.13, *max* = 0.88) and the sample sizes for the specifications range from 21 to 239, with a median of 83. Compared to the first SCA, the second SCA which only had *z*-score as a dependent variable option and linear model as a modelling option, the predictors now account for more variance across model specifications with 45.32%. Followed by residual error (42.26%) and subsets (12.42%). Again, the control variables (gender and age) accounted for 0% of variance across model specifications. The specification choices for the independent variables and subsets are presented in Appendix C to allow for further inspection.

Figure 5.4 provides an alternative way of visualising the SCA results, showing the boxplots which reflect the distribution of effects associated with each parameter across all specifications. Each category (e.g., coefficients, subsets) are given a different colour, and

each boxplot displays the median effect estimate associated a given specification and the associated variance of the effects. The red dots represent extreme/outlier values.

Figure 5.1

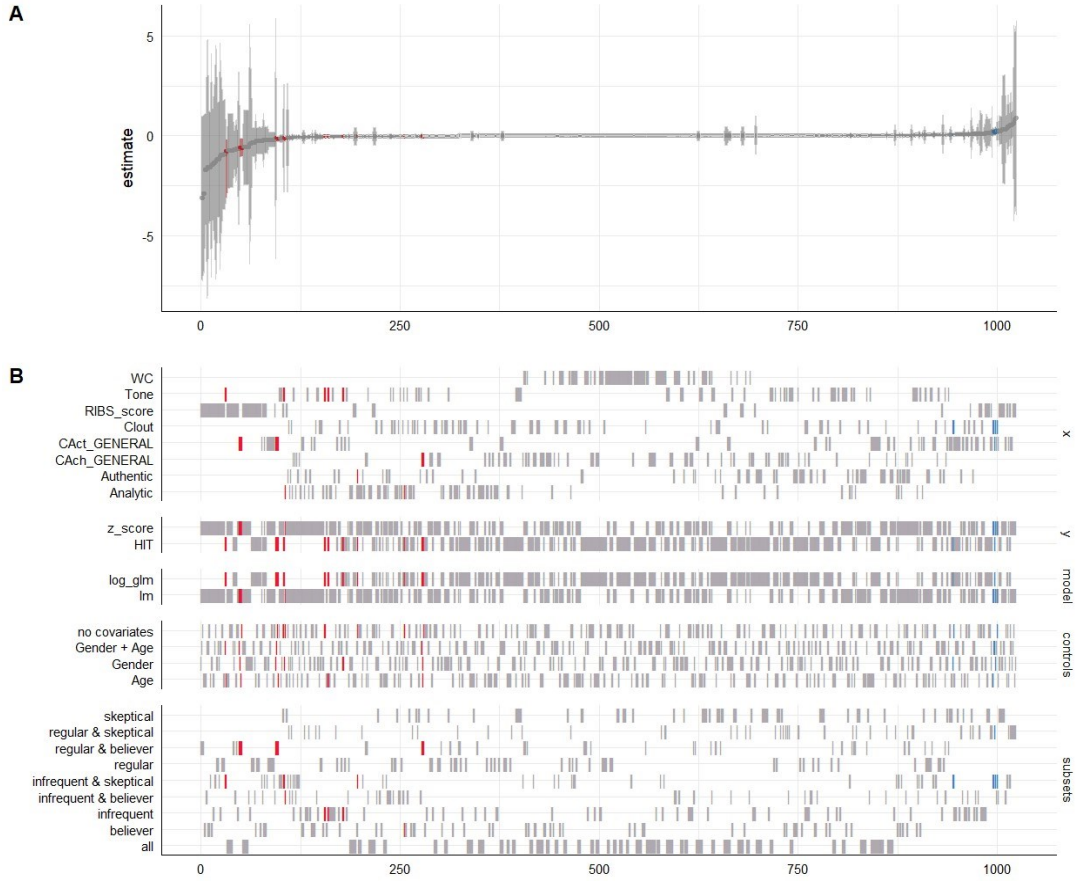
Number of Specification Decisions



Note. Panel A is a boxplot demonstrating the number of options for each variable. Panel B shows a decision tree of all possible combinations, with each node indicating a decision. Total decisions is 1024.

Figure 5.2

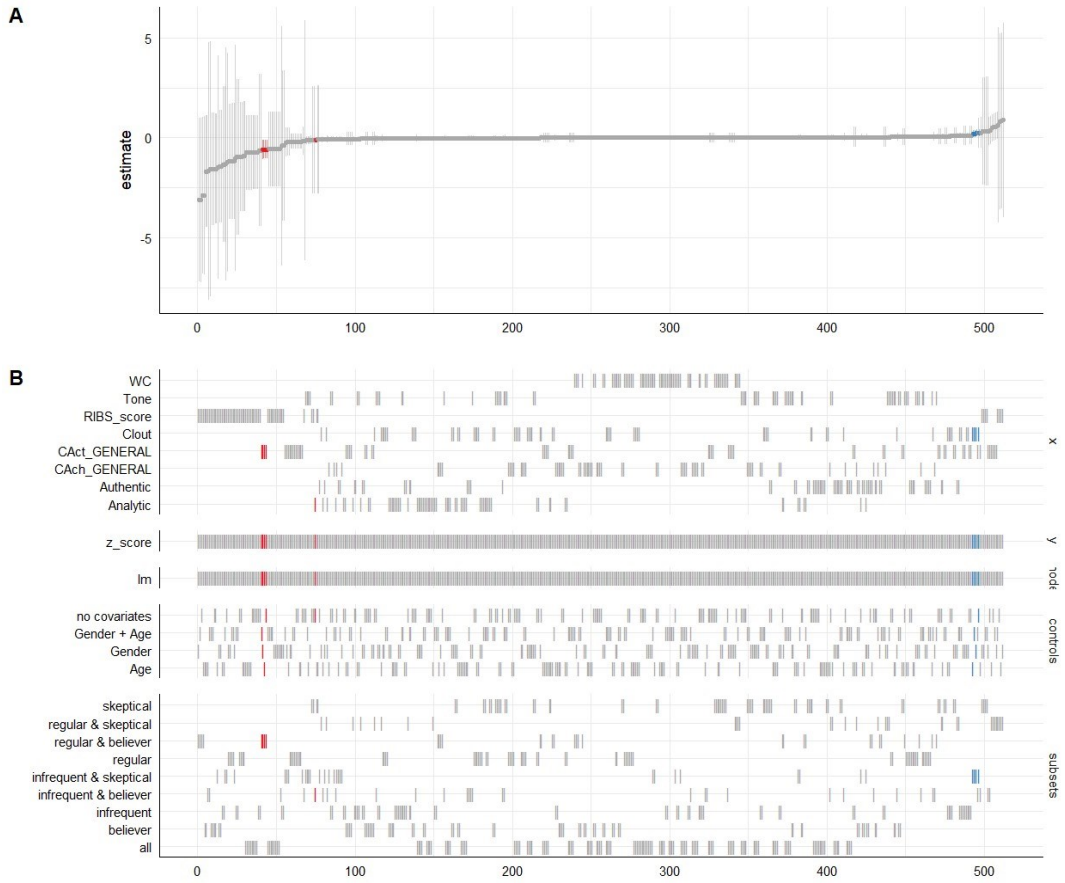
Overall Specification Curve and Specification Decisions with All Variable Options



Note. Red denotes a negative significant effect observed in a given specification, grey denotes non-significant effect, and blue denotes a positive significant effect, all at the 5% significance level.

Figure 5.3

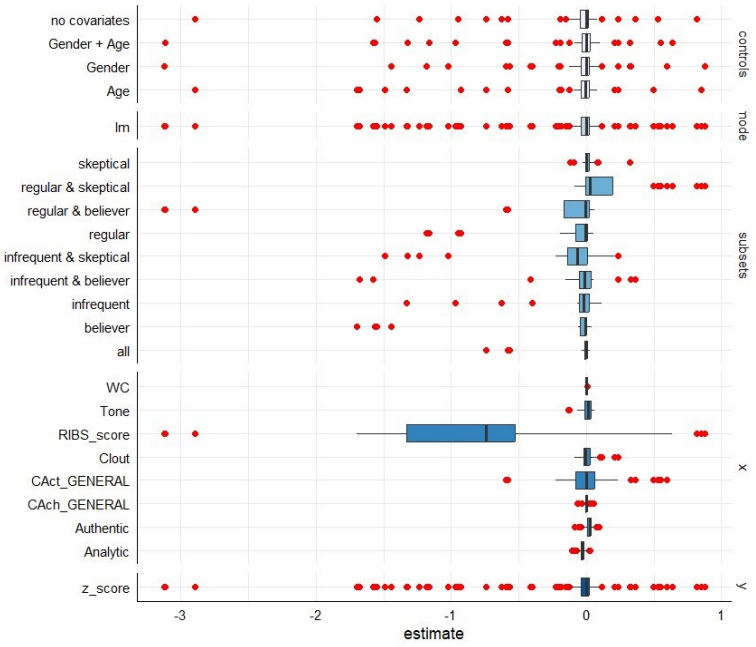
Overall Specification Curve and Specification Decisions with z-scores only



Note. Red denotes a negative significant effect observed in a given specification, grey denotes non-significant effect, and blue denotes a positive significant effect, all at the 5% significance level.

Figure 5.4

Boxplots of Parameter Distributions for z-score Specification Curve Analysis



Note. Specification decisions are grouped by boxplot colour. Red dots represent extreme/outlier values.

6.6 Discussion

As outlined earlier in this chapter, the specification curve analysis was deliberately limited to a single dataset, specific predictors, and grouping variables. Although there are many other justifiable ways to construct and analyse this dataset, even within these constraints, there are between 512 and 1,024 model specifications. This analysis was intended to be exploratory rather than to identify which specifications produce the largest effect sizes. Despite this exploratory nature, the surprising finding is that the vast majority of model specifications produce null effects. What is even more intriguing is that the specifications producing significantly positive results in the second SCA (refer to Figure 6.3) are those that include the Clout mentation variable and participants who are skeptical (i.e., those with below-average ASGS scores) and infrequently practice mental discipline (monthly or less). This is contrary to the belief held in the psi ganzfeld literature, which suggests that higher levels of these traits should correlate with increased session outcome. However, a closer examination of the literature reveals that while the psi ganzfeld community often selects participants based on creativity, mental discipline, thinking styles, paranormal belief, and paranormal experiences, the relationships between these characteristics and session success

are mixed-to-null, except for measures of creativity. Yet, the ICAA measures include in the SCA reported in this chapter (and KPU Study 1074 funding report results) found no association between these creativity measures and session outcome. However, due to participants selected for creativity, a correlation between the creativity variables and session outcome might suffer from range restriction.

A brief overview of previous ganzfeld studies that measured these participant characteristics reveals a mixed pattern of findings (note that this is not an exhaustive literature review). Studies that assessed creativity reported a significant positive relationship between participant creativity and session outcome (Dalton, 1997; Holt et al., 2004; Morris et al., 1993, 2003; Parker et al., 2000; Schlitz & Honorton, 1992). However, studies that included assessments of creative thinking styles, namely the Torrance Tests of Creative Thinking, found no significant relationship between these thinking styles and session outcomes (e.g., Dalton, 1997; Roe et al., 2001; Schlitz & Honorton, 1992). Similarly, studies that used the Australian Sheep-Goat Scale (ASGS) to measure paranormal belief found no significant relationship between ASGS scores and session outcomes (e.g., Milton & Wiseman, 1999; Morris et al., 1993; Parker, 2010; Parker & Westerlund, 1998; Watt et al., 2020). The evidence is mixed in studies that measured previous paranormal or psi experiences and the practice of mental disciplines (e.g., Bierman et al., 1993; Broughton & Alexander, 1997; Honorton, 1997; Honorton & Schechter, 1986; Kanthamani & Broughton, 1994; Marcusson-Clavertz & Cardeña, 2011; Morris et al., 1993; Parker et al., 2000).

This brief summary of characteristics in the psi ganzfeld raise questions about the validity and replicability of psi ganzfeld effects and suggest a more significant concern for researchers in this field. Even when using the same measure, such as the ASGS, there is no replicable evidence of a relationship between paranormal belief and psi ganzfeld task outcomes (e.g., Milton & Wiseman, 1999), yet this measure continues to be examined and included as a guiding characteristic in such studies. This raises the possibility that either the ASGS does not accurately measure paranormal belief, or that paranormal belief itself does not influence psi ganzfeld task performance. If the latter is true, it suggests that psi ganzfeld researchers need to reconsider which participant characteristics are truly relevant to task success. A similar concern is the sheer variety of questionnaires, scales, and items used across psi ganzfeld studies. While not examined in depth in this thesis, a previous review by Goulding and Parker (2001) found a total of 149 different questionnaires, tests, and item sets used across 147 free-

response ESP studies. This inconsistency adds another layer of complexity to understanding and replicating the psi ganzfeld effect.

Relating to earlier discussions, particularly in chapter 4, the psi ganzfeld lacks a set definition or universally agreed-upon characteristics (Milton, 1999; Schmeidler & Edge, 1999). Even in the preregistration for KPU Study 1074, the principal investigator defined their own ganzfeld characteristics. This suggests that each ganzfeld study is unique, shaped by individual researchers' interpretations, participant samples, and experimental teams, along with the use of varying questionnaires and items. Consequently, the findings of these studies may not be as generalisable or replicable as researchers hope. This issue does not point to a fault with any single parapsychologist but highlights a broader, systematic concern within experimental parapsychology, including psi ganzfeld research. The lack of a clear definition for the ganzfeld method calls for a more systematic agreement among researchers, which could be developed through collaborative approaches, such as the Delphi method.

Concerns about experimental parapsychology are not new. For instance, Braude (1992) argued that researchers in this field follow the methodologies of experimental psychology too closely and should instead focus on being "good observers". This perspective is echoed in recent discussions by philosophy of science scholars, who suggest that psychology requires more observation and theory-building work (e.g., Feest, 2024; Lavelle, 2022). Both Feest (2024) and Scheel et al. (2021) advocate for greater emphasis on concept formation and exploratory research, including exploratory experiments and formal modelling, before transitioning to confirmatory research. As Scheel et al. (2021) points out, "...by focusing primarily on confirmatory research and jumping straight to the hypothesis test, psychologists too often neglect the groundwork that is necessary to ensure a sound link between the test and the test theory" (p. 746). To develop better theories, more observational research on spontaneous psi cases and the associated attributes and characteristics is needed (Braude, 1992). Once these foundations are established, the relationship between the test (psi ganzfeld) and the underlying theory will allow for more accurate predictions of observable outcomes (Scheel et al., 2021). While statistical modelling is crucial at the end of this derivation chain and is often highly specific, as Scheel et al. (2021) notes, until the psi ganzfeld has an agreed definition and a testable theory, the current focus on statistical modelling, as demonstrated in this chapter, may be premature.

6.7 Conclusion

Two specification curve analyses were conducted to illustrate the flexibility and control researchers can exert in the statistical analysis of data. The results reveal that most of the possible modelling approaches for data from a precognition psi ganzfeld study produce null effects. This finding suggests that while statistical modelling is a crucial part of the derivation chain between theory and hypothesis testing, psi ganzfeld researchers may face more fundamental concerns (e.g., Braude, 1992; Wiseman, 2010). However, it is more challenging to explore experimental permutations than statistical modelling permutations as the former is much more resource intensive.

Multiverse type analyses can demonstrate the inherent flexibility and ambiguity a researcher is faced when trying to analyse data to answer a hypothesis. By using a contentious topic, this chapter reveals that very few models will produce significant results. By conducting such an analysis, it may provide direction for some researchers, such as which participants to recruit to better understand psi and acknowledge modelling approaches not previously considered. Yet, it also opens the door to researchers selectively using models to produce a result that aligns with their expectations of beliefs.

Chapter 7

Was it something I said? Mentation reports in the psi ganzfeld

7.1 Chapter overview

This chapter provides a concise overview of previous assessments of mentation reports produced by participants in the psi ganzfeld. The chapter summarises the role of introspection, especially within the context of psi ganzfeld. With this context in mind, the main body of the chapter describes mentation reports from precognition studies conducted at the Koestler Parapsychology Unit (KPU; ID no. 1039 and 1074). These mentations were transcribed from audio files and the resulting text files were analysed using the quantitative language model, Linguistic Inquiry and Word Count (LIWC2015). As this is the first instance of quantitative language assessment of mentations, the chapter provides a description of these features and themes identified.

Given the lack of accepted theory in psi research, there is a long history of experimental 'ritual' and 'lab lore' with assumptions that the ganzfeld stimulation produces an altered state of consciousness. Even with a lack of an agreed upon paradigm, participant mentations are routinely collected yet under-explored (Cardeña, 2020; Schmeidler & Edge, 1999). By analysing this data, it will help identify previously unknown facets of the participant experience.

7.2 Previous assessment of the mentation

The mentation is a core feature of the psi ganzfeld paradigm across different modalities (precognition, clairvoyance, and telepathy). Despite ongoing debates regarding core aspects of the ganzfeld (see Chapter 4 for further discussion), such as noise frequency and type (Stanford et al., 1989b; Symmons & Morris, 1997), light colour (Kübel et al., 2021), and other characteristics (Schmeidler & Edge, 1999), the mentation remains a central element. Participants (and sometimes senders) produce mentations by verbally reporting and describing images, sensations, and feelings that come to mind (Wooffitt et al., 2010). The

primary purpose of the mentation is to serve as an *aide-mémoire* for participants during the subsequent judging phase of the psi ganzfeld, framing their assessment of the target clips (Wooffitt et al., 2010). If the ganzfeld is conducive to psi phenomena, the imagery and sensations reported in the mentation may capture anomalous information transfer, aiding participants in selecting the correct target (Wooffitt et al., 2010). But, as one of the guiding principles of the psi ganzfeld task is that it produces an internal attention state or altered state of consciousness (Honorton, 1977; Stanford et al., 1989a), the mentations may shed light on the experience of consciousness (Cardeña & Pekala, 2014).

As discussed in chapter 1, it is often assumed that during a ganzfeld session, participants experience a shift in consciousness, commonly referred to as an altered state of consciousness (ASC). An ASC is defined as a qualitatively different state of consciousness, or a unique pattern of psychological structures, that differs from the ordinary waking state (Cardeña et al., 2015). Research on altered states of consciousness and the ganzfeld began in the late 1970s, with L. W. Braud and Braud (1974), W. G. Braud and Braud (1974) introducing the concept of the psi-conducive state. According to the noise-reduction hypothesis, both external (sensory-perceptual) and internal (emotions, analytical thoughts, etc.) stimuli are reduced during a ganzfeld session, allowing attention to be more available for accessing psi "signals" that might otherwise be overlooked (W. G. Braud, 2002). This suggests that psi is enhanced when participants are in a state of sensory relaxation with minimal influence from ordinary perception (Roney-Dougal, 2015). The noise-reduction hypothesis posits that the ganzfeld reduces both internal and external stimuli, creating a hypnagogiclike state—a transitional state between wakefulness and sleep (W. G. Braud, 2002; Honorton, 1977; Wackermann et al., 2002).

However, some argue that physiological evidence to support the claim that the ganzfeld induces a hypnagogic state is lacking, suggesting instead that it is more similar to the waking state than to sleep onset (Roe, 2009; Wackermann et al., 2002). Nevertheless, some psychophysiological research indicates that imagery induced by the ganzfeld shows remarkable similarities to hypnagogic hallucinations and imagery (Vaitl et al., 2005; Wackermann et al., 2008).

Wooffitt et al. (2010, p. 15) defines the mentation as "...a series of discursive acts through which participants pragmatically address institutional, interpersonal and inferential contingencies of the setting". Wooffitt and colleagues highlight three key aspects of the

mentation. First, the institutional setting: participants are aware of the experimental context in which the mentation is being produced, which may create a pressure to be a "good" participant and report something, even if there is nothing to report (Cardeña, 2004). Second, the interpersonal factor of conveying descriptions of their own experience to the experimenter and potentially to a sender in a different room, may create demand characteristics. Inferentially, verbal reports may have personal relevance, such as belief and memories, but also relate to the task at hand. When people provide introspective descriptions of their inner experience, they perform social actions influenced by the institutional setting and informed by their own conduct and that of the researcher (Wooffitt & Holt, 2011).

Since the early development of the psi ganzfeld, the value of the mentation has been recognised. Honorton (1972) noted the mentation content could be a more sensitive indicator of potential psi information than overall task performance metrics like session *z*-score, hit rate, and binary hit/miss. Stanford et al. (1989a) argued about the virtues of measuring psychological functioning through mentation verbal indicators. The virtues of this approach include: (a) examining the psychological consequences of experimental manipulation, (b) developing indices of subjects' psychological functioning in the setting in which they are testing, and (c) examining the relationships between these factors and ESP-task performance. This allows researchers to get closer to understanding the psychology of the task, rather than relying on personality or demographic variables.

In the 1980s, researchers examined global features of mentations via post-session questionnaires, while others attempted to classify and identify specific characteristics in the verbal reports. Julie Milton's (1986) doctoral thesis attempted to categorise mentations by the nature of their content and experiential qualities and their influence on task outcome. Independent judges coded the mentations and judged the targets based on the mentation content. One judge found a significant negative correlation ($p < .01$, two-tailed) between session *z*-score and a factor composed of participant good mood and pleasantness of the session. Fleeting experiences in the ganzfeld were associated with significantly worse session outcomes than non-fleeting experiences, according to one judge. In a third study, Milton investigated whether target-related mentations could improve target scoring and found no mentation correspondence types were more accurate than others, nor did any type of mentation perform than another. However, Milton stated that any significant findings in her work were inconclusive due to the considerable number of analyses performed on the data.

In a similar vein, Deborah Delanoy (1988) explored various facets of mentation features. Her study examined 15 different types of mentations, such as bizarre, fleeting, and auditory experiences, to determine if any specific response type was significantly associated with psi-hitting. Using external judges and participants to rate the targets, Delanoy's investigation focused on telepathy, analysing both senders' and receivers' mentations. For receivers' mentations, Delanoy (1988) categorised them into five sub-groups: image, duration, clarity, content and miscellaneous. Each mentation was scored based on the presence of these features, and the ranking of picture targets was determined by the rating points assigned by subjects to each picture for each aspect of mentation. Despite these efforts, Delanoy concluded that neither weak nor strong mentation-target correspondences conveyed a significant degree of target-related information. However, one category, 'undeveloped imagery', out of the 15, showed a significantly greater proportion of target-related information.

Like Delanoy, Stanford et al. (1989b) used mentations to measure psychologicalcognitive functioning within the ganzfeld, employing different types and levels of noise, including a silent condition. In their first study, the researchers created four verbal markers:

1. Mean standard deviation of utterance length: This marker was intended to reflect spontaneity;

2. Mean utterance length: This was used to measure arousal, with the belief that lengthy or complex sentences would indicate a reasonable level of arousal;

3. Words per minute: A measure of productivity and arousal. The researchers posited that participants who spoke more frequently had a better chance of mentioning something relevant to the target, but also to the decoys;

4. Proportion of concrete nouns: A high concentration of concrete nouns was thought to indicate a participant who spent most of the session naming a series of objects that came to mind, with little description or commentary

The researchers found that mean utterance length was dependent on a certain minimum level of arousal, whilst words per minute increased and then dropped off steadily to the end of the session. Participants in the noise condition (as compared to the silent) produced longer utterances but there was no evidence that the level of arousal affected the rate of speech.

In the second part of the project, the authors (Stanford et al., 1989a) assessed the internal attention state, which is argued to be facilitated by the psi ganzfeld. This focus on internal attention is a guiding principle of the paradigm's development (see Honorton, 1972, 1985), with the belief that achieving such a state favours psi task performance. Stanford and colleagues found that absorption scales significantly predicted their verbal measure, the quadratic temporal trend of words per minute (QTWPM). The authors suggest that an internal-attention focus in the ganzfeld provides a characteristics verbal "signature" in the form of substantial and sustained involvement with mentation that develops within nine minutes or less, as reflected in the rate of speech. High absorption participants exhibited a temporal trend in which their rate of speech rapidly increased from the first to the second quadrant, was maintained in the second and third quadrants, and then dropped off in the final quadrant. The authors concluded that high absorption participants quickly became engaged in the session, sustained their involvement, and then tapered off towards the end. In contrast, low absorption participants began at the same speaking rate as the high absorption participants and gradually increased their rate of speech, but not to the same extent as the high absorption participants.

Their second verbal measure, the linear trend of proportion of concrete nouns (LTPCN), which measured internal attention, found that people who scored high in absorption tended to show regular increases in the use of concrete nouns as the session progressed.

However, when testing these verbal measures, they did not find supporting relationships between either of their two measures and task performance. They concluded that the premise on which the psi ganzfeld was built—that a psi-favourable internal state enhances task performance—has little support. They also suggested that absorption may not be a significant factor in ganzfeld ESP task performance (Stanford et al., 1989a).

A couple years later, Stanford and Frank (1991) attempted to replicate their previous studies (Stanford et al., 1989a, 1989b). Again, they focused on the rate of speech, as measured by the QTWPM measure, and absorption. In this replication, the authors found no relationship between their verbal marker and internal-attention focus.

Other researchers created their own measures to categorise mentations. Parker and Westerlund (1998) explored repetitive themes in telepathy mentations and hypothesised that more repetitive themes would be positively correlated with session success. Two independent raters evaluated 90 written mentations for repetitive themes. First, they identified the themes

and then counted the frequency of these themes, defined as the highest number of times a given theme reoccurred. The authors acknowledged this measure was insensitive, as multiple themes could be repeated in the same session. Ultimately, they found no relationship between the number of repetitive themes and session outcome.

In a later study, Parker (2006) developed the real-time ganzfeld where mentations were recorded in real-time as the sender viewed the target. The study had two parts: the first was to compare the number of subjective remarkable correspondences between mentations and target/decoy clips. The authors used a 'single utterance' concept to differentiate between segments of correspondences and complete mentations. Out of 20 remarkable correspondences highlighted by the judges, 6 were correct with the target clip, and 14 corresponded to decoys. They concluded that the correspondences between ganzfeld mentation and target content were not that remarkable. In their second part, students rated the "impressiveness" of the 20 correspondences. The hit mentations were given a higher average rating of impressiveness than the decoys, but to a non-significant degree. This led to the conclusion that remarkable correspondences are perhaps due to subjective validation and there is a danger of relying on subjective impressions as validation of paranormal phenomena (Parker, 2006).

At a similar time, Carpenter (2004), created a categorisation of mentations using 364 ganzfeld mentations collected across multiple laboratories, including PRL, FRNM and Dalton's doctoral thesis (Broughton & Alexander, 1997; Dalton, 1997; Honorton et al., 1990). A set of 36 rating scales was employed to assess the approach and quality of the participant's experience. These scales covered various aspects such as image, memory, physical experience, which were then grouped into smaller categories like cognitive aspects, imagery, emotions and colour. The scales were subsequently reduced to nine composite scales to create the *Ganzpred* score. The nine scales were:

1. Positive/neutral experience: Physical or emotional experiences that are pleasant or emotionally neutral;
2. Fluid development: Content or activity of the image that develops over time;
3. Form with achromatic colour;
4. Autonomy: Images with autonomous power or will, sometimes in defiance of the participant's wishes;

5. Cooperative movement: Images showing humans interacting in a non-conflictual manner;
6. Merger/harmony: Images with connotations of positive merger, love or selftranscendence;
7. Integration: Image composed of more than one element, and elements are combined in some way
8. Anxiety: Images with fearful or distressing aspects;
9. Intellectualisation: Images suggesting an intellectualised approach to the task

The authors reported that their *Ganzpred* score correlated with task outcome (session *z*-score). Participants were then divided into three quartiles (high, middle two, low) based upon their *Ganzpred* score. The high-quartile group of *Ganzpred* had 40% first-rank hits, whereas the low-quartile group yielded 25% first-rank hits. Verbal production (number of words) had a negative, significant correlation with session *z*-score ($r = .10$, $p < .05$) whereas positive/neutral experience ($r = .14$, $p < .001$) and merger/harmony ($r = .12$, $p < .001$) were positively correlated with session outcome. These findings suggest that psi hitting is facilitated by positive/neutral experiences rather than uncomfortable or odd experiences. Additionally, artists produced significantly fewer words on average but more independent ideas per transcript, suggesting that people with fewer creative inner resources might compensate by producing extra verbiage and elaboration of relatively few ideas (Carpenter, 2004).

Similarly, Kathy Dalton (1997) recruited participants with creative backgrounds such as writers, actors, musicians, and artists. For the mentation variables of these artists, there was a significant negative correlation between the amount of mentation and session z-score ($r = -.342$, $p < .02$). However, musicians' mentation variables revealed a significant correlation between structured mental activity and z-score ($rho = -.504$, $p < .01$), suggesting that unstructured thought may act as a distraction or inhibitor of the psi process, contrary to Delanoy (1988) findings. Combining the artist and musician groups, the amount of mentation had a non-significant negative relationship ($rho = -.146$).

Dalton's studies involving actors and writers found that the mentation produced by actors was not significant, as indicated by a non-significant z-score ($rho = .006$). However, mentation produced by writers had a significant negative correlation ($rho = -.342$, $p < .02$), indicating that those who produced more content had lower session z-scores. When

combining the results of actors and writers, there was an overall non-significant relationship between the amount of mentation and session z-score (*rho* = -.079).

Overall, the assessment of mentations from psi ganzfeld studies leave little clarity on verbal patterns and trends that may, or may not, be influencing session outcome. Yet, one consistent finding across these studies is that the higher amount of verbal production during the ganzfeld session correlates negatively with session *z*-scores. This suggests that excessive or unstructured mentation may not facilitate psi task performance or excessive mentation is harder to judge. Further, studies by Milton (1986) and Carpenter (2004) suggest a positive correlation between positive or neutral emotional states during the ganzfeld session and *z*-scores.

A caveat to these findings is the multitude of categorisation scales and measures to analyse mentations, complicating understanding of any consistent and reliable findings. Moreover, studies diverge in the form of mentation. Researchers have used different formats, including written reports, audio transcriptions, real-time recording and playback, and thematic analysis of utterances. These methodological differences introduce further complexity and challenges when comparing findings. Regardless of the mixed findings, recent reports by Cardeña (2020) and Stanford (2020) underscore the importance of revisiting these methodologies to gain deeper insights into the role of mentations in ganzfeld experiments.

7.2.1 The issue of introspection

Introspection is the detailed observation of one's lived experience (Petitmengin & Bitbol, 2009) and is at the core of understanding how participants perceive and articulate their experiences during the ganzfeld session. The next part of this chapter will explore the role of introspection, its value and implications, and research into introspection in ganzfeld studies.

General issues of introspective methods

Petitmengin and Bitbol (2009) outline many common arguments against introspection, as it is often argued that reporting on one's lived experience is either impossible or introduces irreducible distortion. Other pitfalls include observational, temporal, interpretative and verbal distortions. The first argument, known as the 'impossible split' and

observational distortion, questions how one can distance themselves from an experience whilst trying to report it accurately. This suggests a gap between the stimulus and subsequent report. However, introspection is not merely about observing and describing stimuli; it is about observing and describing one's own *experience* of those stimuli. Petitmengin and Bitbol (2009) argue that becoming aware of one's own experience is not distancing, but rather reducing the distance and coming closer to the experience. However, in the ganzfeld setting, this dual focus on describing the target and one's own experience of the target blurs this line. But Cardeña and Pekala (2014) argue that the issue is not whether introspective methods affect the content of the conscious experience, but whether they distort it enough to invalidate the obtained data.

Temporal distortion is another challenge; reporting an experience is inherently different from the state of having the experience itself. The act of evocation, or bringing the experience into reflective consciousness, allows dimensions of the experience to appear that may have gone unnoticed initially. This creates a new experience in the present moment of reporting, which can differ from the original one. Petitmengin and Bitbol (2009) argue that we cannot live an experience 'in the past' as there is no other experience other than the present, thus, in this evocation state we live a new experience. This is less of a concern for the psi ganzfeld task as participants are asked to report their experiences live, yet there will remain a time lag between the experience itself and describing it verbally.

Interpretative distortion raises concerns about misinterpreting one's own experiences due to preconceptions or beliefs which can distort the introspective process (Bitbol & Petitmengin, 2013; Cardeña & Pekala, 2014; Petitmengin & Bitbol, 2009). This is particularly relevant in the ganzfeld setting, where participants are often unfamiliar with the procedure and may have strong preconceived notions about psi phenomena. However, some research laboratories attempt to overcome this unfamiliarity by giving free-response judging guidelines to the participants (Delanoy et al., 2004; Watt et al., 2020). Nonetheless, these factors can lead to altered or socially desirable descriptions of their experiences (Cardeña, 2004). Yet, Petitmengin and Bitbol (2009) argue that the more individuals engage with their experience, the simpler, more direct, and concrete their descriptions become, indicating a genuine connection to their experience. Suggesting that the more engaged an individual, the less chance for bias and expectation to influence their introspection.

Verbal distortion states that words can transform and potentially distort the original experience since words can only point to, but not fully encapsulate, the experience (Cardeña & Pekala, 2014). However, Petitmengin and Bitbol (2009) counter by arguing that "in themselves, words are empty, they only become meaningful through the gesture that relates them to experience" (p.390). Thus, the focus should be on the quality of the descriptive *process*, rather than the absolute validity of the descriptions.

Finally, criticisms of introspection within scientific methodology centre on the private and singular nature of personal experiences, which are neither easily verifiable nor falsifiable (Bitbol & Petitmengin, 2013; Cardeña & Pekala, 2014; Petitmengin & Bitbol, 2009). Despite these criticisms, introspection remains a crucial method for accessing and understanding the internal experiences that form the basis of mentation reports in ganzfeld research.

Introspection in the ganzfeld

Contemporary methods for introspection in ganzfeld research have been influenced by the 'think-aloud' protocol associated with Ericsson and Simon (1980), as discussed by Cardeña and Pekala (2014) and Wooffitt and Holt (2011). This protocol involves participants vocalising their thoughts while performing a task, aiming to capture their inner experience in real-time. Ericsson and Simon (1980) argue that the think-aloud protocol offers the closest possible match between inner experience and verbal report by *limiting* introspection to the content of the experience, thereby reducing issues of interpretation and memory distortion (Cardeña & Pekala, 2014).

The ganzfeld paradigm presents a unique case for the application of the thinkaloud protocol. In this setting, the participant is alone, unlike normal conversation contexts that involve turn-taking. The relaxed environment and singular task of providing a monologue for approximately 30 minutes to an overhearing, but nonparticipating, experimenter creates a unique scenario. Participants understand the experiment's purpose and conditions, which influences their imagery reports (Wooffitt & Holt, 2011; Wooffitt et al., 2010).

Although there is discourse about the validity of introspection and first-person reports, there is a growing awareness of the value in studying consciousness (Cardeña & Pekala, 2014). Wooffitt et al. (2010) analysed mentations from the Koestler Parapsychology Unit at the University of Edinburgh and observed that mentation reports produced in the psi ganzfeld are composed of two phenomena: talk and silence. The talk can range from minimal imagery

reports with one or two words, to more developed descriptions, including references to people, cultural figures, activities, sensations, environments, and even the immediate environment of the experiment itself, such as the white noise. These images often defy categorisation, with some being described once whilst others are recurring. Some imagery is presented as unambiguous, while other reports suggest the participant's doubt about their experience. Silence, which is notably longer in ganzfeld mentations compared to everyday conversations, plays a crucial role. The absence of talk manages the participant's contributions, with silences defining discrete images or sensations.

In addition, Wooffitt et al. (2010) identified two attention markers in ganzfeld mentations: visual and cognitive markers. The use of first-person narrative and verb tenses like "see" and "seeing" indicates that participants experience visual imagery as 'live' from the beginning of the mentation period. Additionally, cognitive markers such as "I'm thinking" frequently appeared in the reports, either expressing explicit doubt about the experience or confidence in describing it. These markers may also subtly frame the participant's experiences as having non-parapsychological origins.

Overall, think-aloud protocols, such as the ganzfeld mentation, offer rich insights into internal experiences by capturing real-time thoughts and feelings. However, introspection can lead to observational, temporal, interpretative, and verbal distortions, potentially altering or misrepresenting the original experience. Despite these limitations, the benefits of gaining access to otherwise inaccessible cognitive and emotional process make introspective methods valuable.

As mentations are routinely gathered during the psi ganzfeld, they offer a vast amount of data and information that has not been fully examined with modern methods. By applying quantitative language models, researchers can potentially uncover new insights and understandings that were previously unattainable. The awareness of the value of the mentation has recently been highlighted in the parapsychology literature. Cardeña (2020) called for further research into the data, using both qualitative and quantitative methods. Likewise, Stanford (2020) directly called for researchers to use computerised transcript analysis to uncover any previously unattainable patterns in the verbal reports. Thus, by using a quantitative approach, we enhance our ability to explore the psi hypothesis, any associated language patterns associated with the task, new insights into states of consciousness and experiences of participants.

7.3 Quantitative language analysis of KPU studies 1039 and 1074

The two datasets originate from precognition ganzfeld studies conducted at the University of Edinburgh. Both studies share methodological similarities, including the use of the same study procedure, experimental software and random number generation. Additionally, both studies were conducted by Edinburgh university students, with the primary demographic of participants being Edinburgh university students as well. Given these methodological consistencies, except for minor deviations discussed below, the two datasets have been combined for analysis.

7.3.1 KPU Study 1039

Koestler Parapsychology Unit study 1039 was conducted as part of a final year psychology undergraduate dissertation project from September 2017 - March 2018. Three undergraduate students (one of which was myself) recruited participants and collected the data, with each student collecting 20 sessions each. The study used a precognition paradigm and recruited 60 (unpaid) volunteers to contribute to one session each at the ganzfeld laboratory at the University of Edinburgh. Participants were screened and selected, based on previous literature (see Chapter 1), who selfrated their creativity at 3 or more (out of 5), or practised a mental discipline, or had a prior paranormal experience and/or belief in ESP and PK. Those under the age of 18 and/or had a current mental disorder were excluded. The study produced significant, positive results in support of the psi hypothesis with a 37% hit rate (22/60), compared to the mean chance expectation of 25% (exact binomial $z = 1.94$, $p = .03$, one-tailed). Full details on study recruitment, methodology and planned analyses can be found at https://edin.ac/3WnV6QH and published results are available at https://edin.ac/4d4PFfk.

The participant mentations were audio recorded from the start of the relaxation tape (9 minutes long) through the subsequent white noise phase (25 minutes long) and ended after the mentation review phase. Participants were instructed not to start giving their mentation until the relaxation tape had ended and the white noise had begun. The mentation review phase is after the ganzfeld stimulation but prior to the participant judging the targets. The experimenter in the room was responsible for starting and stopping the audio recorder. Additionally, the experimenter made mentation notes, which the experimenter reviewed with the participant, before the judging phase of the session.

7.3.2 KPU Study 1074

KPU study 1074 was a large scale precognition study, using the same software, design and screening procedure as study 1039. Study 1074 was collected in three waves with a total of 240 (unpaid) volunteers contributing one session each at the ganzfeld laboratory at the University of Edinburgh. The study ran from September 2022 to April 2024. Wave 1 (September 2022 - March 2023) and Wave 3 (September 2023 to March 2024) were collected by two sets of psychology undergraduate dissertation students (3 experimenters per wave), whilst the second wave used larger team of researchers (N = 15) from summer 2023 to spring 2024. Participants were screened and selected, like study 1039. For this study, however, participants had to meet the self-rated creativity criterion (3 or more out of 5) plus at least one of the other three (practice of a mental discipline, prior paranormal experience and/or belief in ESP and PK). This study produced a significant positive result in supporting the psi hypothesis with a 32% hit rate (72/240), compared to the mean chance expectation of 25% (exact binomial z = 1.71, p = .043, one-tailed). Full details on the study recruitment, methodology and planned analyses can be found at https://edin.ac/4eYch2R.

The participant mentations were audio recorded from the start of the relaxation tape (9 minutes long) through the subsequent white noise phase (25 minutes long) and ended when the participant acknowledged the white noise had ceased, before the judging phase. This study did not have a formal mentation review period like study 1039. The experimenter in the room was responsible for starting and stopping the audio recorder. Additionally, the experimenter made mentation notes, which the participant had the option to review independently before the judging phase of the session.

7.4 Method

7.4.1 Mentation transcription

The mentations from KPU study 1039 were transcribed by research assistants, who did not collect the data, and manually listened to each audio file and transcribed the files into Word documents in 2018. They transcribed the ganzfeld mentation and the mentation review period. Out of the 60 session, 54 sessions were transcribed. Missing transcriptions were either due to participants not providing a verbal report, audio recorder malfunction,

participants falling asleep during the session or the participant speaking in a language other than English.

The mentations for KPU study 1074 Wave 1 and 2 were transcribed by two undergraduate students conducting their final year dissertations in 2023-24. The first two waves were transcribed using Microsoft Office 365 transcribe tool in Word. The students listened to each recording via Microsoft Word and edited any incorrect auto transcription. Wave 1 and 2 mentations were saved in Word documents. Wave 3 was transcribed by myself, using the University of Edinburgh's Media Hopper Create during the summer 2024. For this, I uploaded the audio files to the software, requested captions from the service provider, and listened to each file to edit the generated captions where necessary. I then downloaded these captions as text files. A total of 197 out of the 240 session mentations were transcribed. Sessions without transcriptions were either due to participants not providing a verbal report, participants falling asleep during the session, audio recorder malfunctions, or participants speaking in a language other than English. A significant proportion of participants in Wave 3 were Chinese/Hong Kongese, reflecting psychology student demographics, and numerous of these participants spoke in Mandarin/Cantonese for their entirety of their mentation. As neither I nor the other students involved in transcription spoke a language other than English, these mentations were not transcribed. Additionally, translating the mentations could introduce subtle changes that might make the translated versions different from the verbatim mentations.

Thus, the total number of mentations is 251 out of 300 sessions across the two KPU studies. 54 mentations are from study 1039 and 197 from study 1074. The mentation section, not the mentation review period, was used from study 1039.

Linguistic Inquiry and Word Count 2015

The mentation text files were analysed using Linguistic Inquiry and Word Count (LIWC) 2015 software (Pennebaker, Boyd, et al., 2015). LIWC is a closed-vocabulary, lexicon-based tool that is suitable for psychologists and other researchers with little or no background in data science (Dudău & Sava, 2021; Kern et al., 2016). LIWC consists of an internal dictionary and software to provide automated, tokenisation and word counting (Boyd & Schwartz, 2021). LIWC2015 includes approximately 90 analytical dimensions, most of which capture specific features of grammar and vocabulary (e.g., adjectives, pronouns, punctuation usage), with others relate to time, emotion, and cognitive processes. Words

contained in texts that are read by LIWC2015 are called target words and are categorised by the software according to its dictionary, which organises words into groups that tap a particular domain (e.g., negative emotion words). The default dictionary is composed of nearly 6,400 words, word stems (e.g., hungr* allows for any target word that matches to first five letters to be counted an ingestion word, for example) and select emoticons–punctuation marks, letters, and numbers to create pictorial items to represent an emotion. Table C.1 provides a comprehensive list of all LIWC2015 dictionary categories, scales, sample scale words and associated word counts.

The training sets for LIWC2015 were personal blogs, expressive writing by university students, English novels written between 1660 and 2008, *New York Times* articles between January and July 2014, natural speech recorded during people's daily lives and Twitter posts. These training sets were produced by a combined authorship of over 80,000 writers and speakers, totalling 231 million words (Pennebaker, Boyd, et al., 2015).

LIWC2015 also has four summary variables with each variable based on algorithms created by the Pennebaker Lab. Each summary variable has a 100-point scale from 0-100 (Pennebaker, Booth, et al., 2015). The variables are:

- *Analytical thinking* - a high number represents formal, logical and hierarchical thinking, whilst a lower number is more informal, personable, here-and-now and narrative thinking.
- *Clout* - a high number represents that the author is speaking with confidence and a position of high expertise whilst a low number suggests filtering, tentative and even anxiety.
- *Authentic* - higher numbers represent honesty, personal and disclosure, while lower numbers suggest a guarded, distanced form of discourse.
- *Emotional tone* - a high number (over 50) suggests a positive, upbeat style and low number suggests anxiety, sadness or hostility. Ambivalence or lack of emotions are numbers around 50.

This analysis represents the first application of quantitative language software to process ganzfeld mentation reports, and as such, all possible categories were examined to identify trends and patterns, except the punctuation and words per sentence categories. Since

the mentations are verbal rather than written reports, punctuation is less relevant, and all categories except punctuation were analysed to explore potential trends and patterns. Words per sentence was not included as that uses periods, exclamation points and question marks as end-of-sentence markers. As highlighted by Wooffitt and Holt (2011), participants often exhibit distinct patterns, such as alternating statements and silences. However, this analysis does not account for these pauses, treating the text as a continuous section without time markers for silences.

This chapter provides a descriptive analysis to characterise the mentation content of two precognition studies. As this is the first quantitative language analysis of ganzfeld mentations, no specific hypotheses about trends or patterns in the mentation reports are proposed. Nonetheless, the language analysis will hopefully provide a description of what the participants are experiencing during the ganzfeld process. Exploratory correlational analyses are conducted to investigate potential relationships suggested by previous research:

1. Is there a relationship between mentation word count and session *z*-score?
2. Does the emotional tone of the mentation predict session *z*-score?

Links between creative or artistic types and session outcomes are not analysed in this chapter because KPU Study 1039 did not use standardised questionnaires to assess participants' artistic activity and achievement. The mentation text files were analysed using LIWC2015, and the results were imported into JASP statistical software (JASP Team, 2024) to produce descriptive statistics, calculate 95% confidence intervals for the mean using bootstrapping with 1,000 samples, and perform simple linear regression.

7.5 Results

7.5.1 Participant demographics

Out of the 300 participants across the two studies, 251 mentations were transcribed. Among these 251 participants, 177 were female, 68 were male, 5 identified as other (non-binary/third gender/prefer not to say), and one participant's gender was missing. The average age of the participants was 30 years (SD = 14.71), and they had an average Australian Sheep-

Goat Scale (ASGS) score of 18.32 (*SD* = 8.35), indicating that the sample was neither strongly believing nor disbelieving in the paranormal. Mann-Whitney U tests indicated there were no significant differences in ASGS scores and age between participants whose mentations were transcribed and whose were not, with $U = 5647.00$, $p = .39$ for ASGS scores, and $U = 6065.50$, $p = .91$ for age.

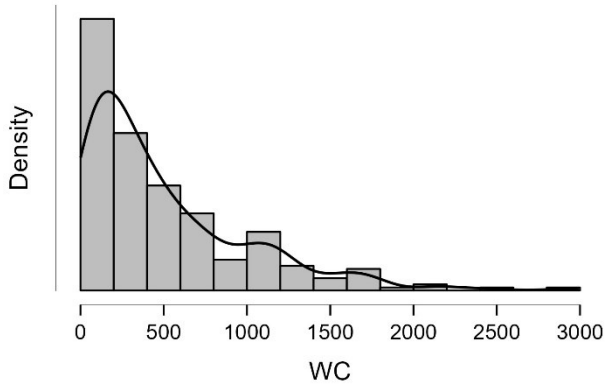
7.5.2 Descriptive Statistics

All Transcribed Mentations

The average word count for the 251 transcribed mentations was 505 words, 95% CI [44.5, 571.37]. Measures of variance illustrate substantial variability among participants, with word counts ranging from 2 to 2942 words. As shown in Figure 7.1, the mentation word count is right skewed, with most participants producing verbal reports less than 500 words.

Figure 7.1

Histogram of Word Count for all 251 Transcribed Mentations



Regarding the four summary variables, Analytic, Clout, Authentic and Tone, the analysis revealed that participants exhibited analytical thinking, showing a moderate-to-high level of logical thinking (*MAnalytic* = 68.66, 95% CI [65.12, 72.11]) during the ganzfeld stimulation. As shown in Figure 7.2a, the negatively skewed histogram reiterates participants using a high number of analytical verbal markers. The Clout variable, however, indicated that participants spoke without confidence and were tentative or anxious in their verbal reports (*MClout* = 32.3, 95% CI [29.94, 34.76]), this is also shown in Figure 7.2c. Despite this, participants were, on average, honest and unguarded when giving these reports (*MAuthentic* = 69.62, 95% CI [66.00, 73.20]). Like the distribution for Analytic, participants show a high usage of

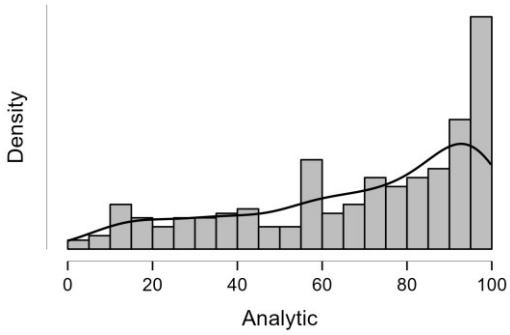
Authentic words, see Figure 7.2b. In terms of emotional tone, the average participant conveyed ambivalence or even feelings of anxiety or sadness during the ganzfeld session (*MTone* = 41.75, 95% CI [38.74, 44.65]). Figure 7.2d shows a more uniform distribution but a noticeable peak in the lower end of the scale, highlighting the overall negative emotional tone of the mentations. Summary statistics for all dimensions of LIWC2015 for all 251 transcribed mentations are provided in Appendix D.

A simple linear regression was fitted to assess if mentation word count influence session outcome but was non-significant (b = -0.00, SE = 0.00, $t(246)$ = -0.14, p = 0.89). A second simple linear regression was fitted to examine if emotional tone of the verbal reports predicted session outcome and was also non-significant (b = 0.01, SE = 0.02, $t(246)$ = 0.55, p = 0.58).

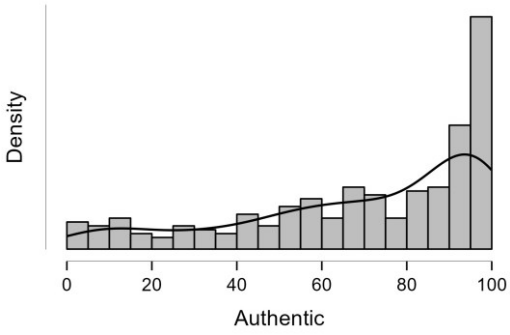
Figure 7.2

Histograms for the Four Summary LIWC2015 Variables

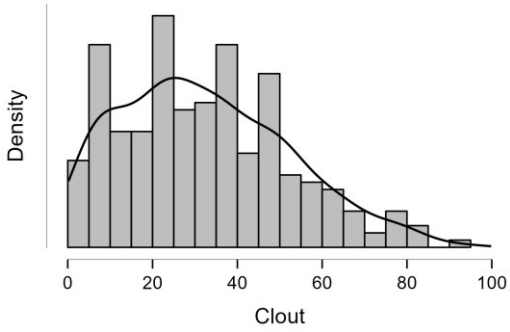
(a) Analytic



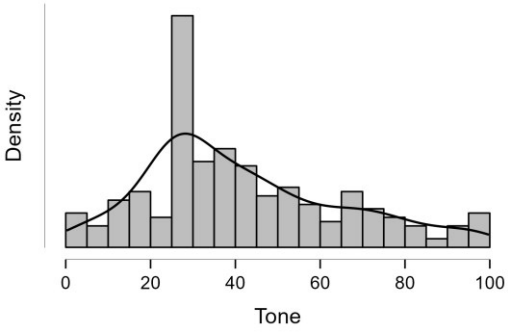
(b) Authentic



(c) Clout



(d) Tone

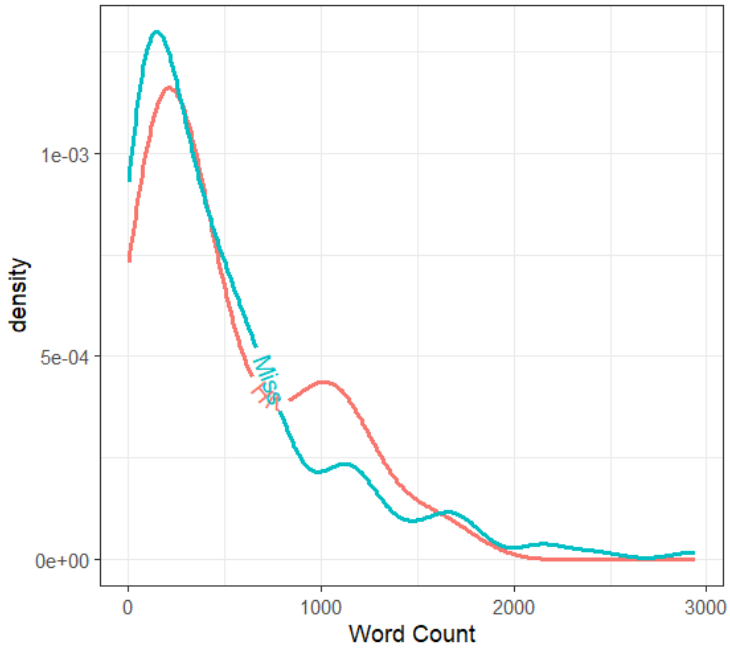


Hitters versus Missers

Out of the 300 total participants across the two studies, 206 recorded a miss and 170 of these sessions were successfully transcribed. Of the 94 (out of 300) participants who recorded a hit, mentation reports from 81 of these participants were transcribed. The average number of words per mentation 523.22, 95% CI [437.29, 619.50] for hitters and 496.50 words for missers (95% CI [412.74, 577.57]) for missers. There was no statistically significant difference between these two means.

Figure 7.3

Density Plot of Word Count by Session Outcome

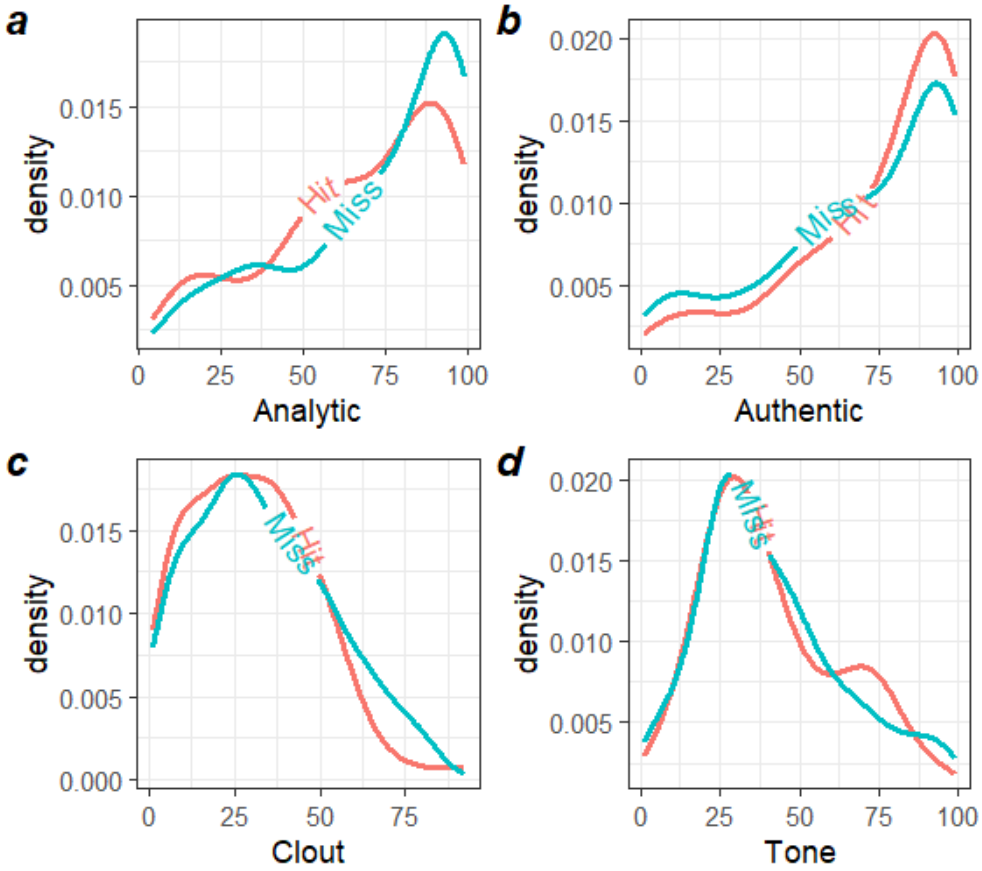


With respect to the four summary LIWC2015 summary variables, hitters and missers used similar proportions of words captured by these metrics (see Figure 7.4). On average, hitters scored less on the Analytic dimension (M = 65.48, 95% CI [59.49, 71.36]) compared to missers (M = 70.18, 95% CI [66.01, 74.45]). For the Clout variable, hitters had a mean score of 30.29 (95% CI [26.31, 34.44]), which is comparable to that of missers (M = 33.27, 95% CI [30.27, 36.05]). For Authenticity, hitters exhibited slightly higher average scores (M = 73.36, 95% CI [67.36, 79.30]) than missers (M = 67.83, 95% [CI 63.64, 72.09]). Both groups showed similar Tone scores, indicating a generally ambivalent or slightly negative experiences during the ganzfeld stimulation: hitters had a mean Tone score of 42.07 (95% CI [37.01, 46.90]) and missers 41.60 (95% CI [38.24, 44.80]). No significant differences were

found between hitters and missers across any of the four LIWC2015 summary variables. Full descriptive statistics for all LIWC2015 variables are reported in Appendix D.

Figure 7.4

Density Plot for the Four Summary LIWC2015 Variables by Session Outcome



7.5.3 Exploratory Analysis

Given the large number of LIWC2015 variables and the exploratory nature of the analysis, linear and logistic regression models were used to identify potential predictors of session outcomes based on participants' mentation reports. Two separate models were developed to reflect the two outcome variables: session *z*-score and binary hit rate. Both models employed stepwise selection and bootstrapping with 1,000 replicates.

Session z-score

Beginning with all LIWC2015 variables, stepwise selection identified a best-fitting model comprising four predictors: *shehe* (third-person singular pronouns), *affiliation, differ* and *home.* Across all five top-performing models (see Table 7.1), higher usage of *shehe*

words was associated with lower session *z*-scores. In contrast, greater use of *affiliation* (e.g., ally, friend, social), *differ* (e.g., hasn't, but, else), and *home* (e.g., kitchen, landlord) words consistently predicted higher session *z*-scores. Notably, the presence of *affiliation* in a model was associated with an increase in session *z*-score of at least one standard deviation. Full model statistics and fit metrics are presented in Appendix D.

Table 7.1

Table of Model Bootstrap Coefficients for z-score as outcome

Model	Unstandardised coefficient		Bias	*SE*	*p**
1	(Intercept)	1.165	-0.021	0.564	0.037
2	**(Intercept)**	**1.658**	**-0.034**	**0.611**	**0.007**
	shehe	**-1.306**	**-0.021**	**0.636**	**0.045**
3	(Intercept)	0.865	0.006	0.774	0.286
	shehe	-1.431	-0.029	0.649	0.025
	affiliation	1.318	0.005	0.619	0.031
4	**(Intercept)**	**-0.883**	**0.007**	**1.073**	**0.400**
	shehe	**-1.478**	**-0.035**	**0.661**	**0.017**
	affiliation	**1.525**	**0.040**	**0.601**	**0.014**
	differ	**0.568**	**0.001**	**0.265**	**0.029**
5	(Intercept)	-2.186	-0.058	1.148	0.090
	shehe	-1.510	-0.045	0.677	0.033
	affiliation	1.535	0.034	0.584	0.011
	differ	0.757	0.014	0.256	< .001
	home	0.877	0.008	0.339	0.023

Note. Bootstrapping based on 1000 replicates. Coefficient estimate is based on the median of the bootstrap distribution. * Bias corrected accelerated.

Further examination of the four LIWC2015 variables identified in the session *z*-score model suggests that the negative association between *shehe* usage and session *z*-score is driven by a small subset of participants who recorded a miss while using relatively high frequencies of these pronouns (see Appendix D).

Binary hit rate

For the binary outcome (hit or miss), stepwise model selection began with all LIWC2015 variables included. Across six models, give variables emerged as recurrent predictors: *differ, home, affiliation, leisure*, and *shehe.* The best-fitting model (Model 4; see Table 6.2), included *differ, home* and *affiliation.* As in the session *z*-score model, greater usage of these three word categories was associated with increased odds of a hit. In models that included *shehe* and *leisure* (e.g., cook, chat, movie), higher usage of these categories

predicted a greater likelihood of a miss. Full model statistics and fit indices are presented in Appendix D.

Table 7.2

Table of Model Bootstrap Coefficients with Binary Hit Rate as outcome

Model	Parameter	Estimate	Bias	*SE*	OR
1	(Intercept)	-0.741	-0.006	0.135	0.476
2	**(Intercept)**	**-1.229**	**-0.020**	**0.262**	**0.293**
	differ	**0.160**	**0.003**	**0.067**	**1.173**
3	(Intercept)	-1.567	-0.022	0.296	0.209
	differ	0.208	0.003	0.071	1.231
	home	0.213	0.006	0.109	1.237
4	**(Intercept)**	**-1.846**	**-0.016**	**0.349**	**0.158**
	differ	**0.231**	**0.002**	**0.077**	**1.259**
	home	**0.213**	**-0.008**	**0.115**	**1.237**
	affiliation	**0.297**	**0.005**	**0.152**	**1.345**
5	(Intercept)	-1.556	-0.009	0.387	0.211
	differ	0.199	0.004	0.077	1.221
	home	0.225	0.003	0.116	1.253
	affiliation	0.395	0.025	0.180	1.484
	leisure	-0.146	-0.020	0.114	0.864
6	**(Intercept)**	**-1.477**	**-0.024**	**0.392**	**0.228**
	differ	**0.201**	**0.005**	**0.080**	**1.222**
	home	**0.229**	**0.005**	**0.115**	**1.258**
	affiliation	**0.440**	**0.028**	**0.176**	**1.552**
	leisure	**-0.150**	**-0.011**	**0.105**	**0.861**
	shehe	**-0.302**	**-0.039**	**0.216**	**0.739**

Note. Bootstrapping based on 1000 successful replicates. Coefficient estimate is based on the median of the bootstrap distribution.

7.6 Discussion

In examining mentation reports from the psi ganzfeld using quantitative language analysis for the first time, there is no comparison or baseline group to contextualise the observed patterns. However, the Clout and Authentic summary variables demonstrate participants' honesty in their reports, despite expressing tentative and uncertain feelings about their experiences. This may help to answer Cardeña and Pekala (2014)'s question about the validity of introspection reports in the ganzfeld. The high levels of honesty suggest the mentation reports are not influenced by observational distortion enough to invalidate the data. Rather, the mentation reports exhibit accurate depictions of the ganzfeld *process* participants experience (Petitmengin & Bitbol, 2009).

Contrary to the findings of Milton (1986) and Carpenter (2004), the mentation reports from KPU 1039 and 1074 found that participants had ambivalent or negative experiences of the psi ganzfeld, yet both studies had significantly more hits than chance expectation. Further, the linear regression found no influence of emotional tone on session outcome. Both hitters and missers, on average, had a similar emotional experience, with some hitters having a more pleasant time but there was no significant difference in the emotional tone of those who recorded a hit or a miss. The feeling of ambivalence may support Carpenter (2004)'s findings that neutral experiences produce successful sessions. But, the high levels of tentativeness and cognitive processing suggest that participants in general found the ganzfeld an odd experience and were trying to process it as it happens, counter to Carpenter (2004)'s claim. However, both study samples were predominantly ganzfeld novices and the mentation characteristics may differ if participants had a familiarisation session or had experienced the ganzfeld before.

The high frequency of first-person pronouns and present time focus indicates active participant engagement with the task and their ongoing experiences. The prevalence of analytic thinking suggests a formal and logical approach to describing unfolding experiences, potentially driven by participants' efforts to comprehend novel psi ganzfeld tasks they are unfamiliar with. This trend is reinforced by the increased use of cognitive processing words, particularly those indicating insight and tentativeness, reflecting participants' navigation through unfamiliar sensations and their attempt to articulate and make sense of their experiences in real-time. Both of these findings support Wooffitt et al. (2010)'s findings about cognitive markers: participants imply caution about their imagery and doubt about the relevance of what they are reporting with the experimental task.

Further, the high usage of analytical thinking words contradicts the long-held belief that psi-conducive environments, like the ganzfeld, reduces internal stimuli, such as analytic thoughts (W. G. Braud, 2002). Yet, both KPU studies 1039 and 1074 produced hit rates significantly greater than mean chance expectation, yet the participants were highly analytical in their verbal reports. Although participants who scored a hit demonstrated less analytical thinking than missers, the difference was not statistically significant.

Visual experiences dominate the mentations, aligning with the task's visual stimuli and instructions to keep eyes open during the session, supporting the results of Wooffitt et al. (2010)'s analysis of earlier KPU studies. The frequent use of spatial words further illustrates participants' efforts to spatially orient their visual imagery and describe its arrangement within their perceptual field to the experimenter in the room. These findings collectively underscore the utility of mentations as rich sources of introspective data, shedding light on how participants perceive and articulate their experiences during psi ganzfeld experiments.

For this analysis, the entire mentation report was analysed, rather than breaking into utterances, or 'chunking' as recommended by Westerlund et al. (2006).. Using the whole of the mentation report, we can identify broader trends and patterns similar to earlier categorisation work, whereas chunking would not capture these patterns and may be more susceptible to filtration by the participant.

The exploratory analyses suggest that participants who used more *differentiation* and *affiliation* words were more likely to record higher session *z*-scores or a hit. This pattern may reflect a greater degree of engagement or clarity, particularly in the case of *differentiation* words (e.g., wasn't, hasn't) which could indicate participants' efforts to define or distinguish features of their subjective experience during the ganzfeld stimulation. Conversely, increased usage of third-person singular pronouns (she/he), *leisure* and *home* words may reflect a lack of task engagement, as these categories could relate to content or memories external to the study task. However, it is possible that such words reflect participants' attempts to describe mental imagery (e.g., "I can see a woman in the kitchen, she is cooking"), which could be relevant to the task. The underlying mechanism linking these linguistic features to session outcome remain unclear, but this analysis provides novel insight into potential cognitive-linguistic correlates of psi ganzfeld performance.

Notably, the LIWC2015 summary variables offered litter differentiation between hit and miss sessions. While participants who recorded a hit tended to be slightly less analytical, more authentic, and used language with a more positive emotional tone—these effects were not statistically significant.

7.6.1 Limitations

One limitation of using quantitative language methods, like LIWC2015, is the training data the model was given. The dictionaries employed by LIWC2015 were developed using testing sets that include personal blogs, expressive writing by university students and *New York Times* articles etc. As highlighted by Wooffitt and Holt (2011), the psi ganzfeld is a unique paradigm that does not adhere to regular social conversational norms, such as turn-taking. A study more closely related to parapsychology research is Bulkeley and Graves (2018) analysis of a dream report repository. They found that dream reports contained a higher usage of first-person singular pronouns compared to the LIWC2015 repository, indicating a more introspective and personal nature. Additionally, the Authentic summary measure, which assess the filtering and genuineness of the speaker, was higher in dream reports than in the LIWC2015 testing sets. Without a baseline for ganzfeld mentations in quantitative language analyses, it is evident that these subjective, individual reports, such as dreams or ganzfeld mentations, differ from the testing sets used by LIWC2015. This suggests that while LIWC provides valuable insights, its dictionaries might not fully capture the unique characteristics of ganzfeld mentations, warranting a cautious interpretation of the results.

As noted earlier, silences are a key part of the mentation process, as highlighted by Wooffitt and Holt (2011). This important aspect was not captured in the transcriptions analysed here. Silence does not imply that nothing is happening or that it is unworthy of assessment. Instead, it may indicate that the participant finds it difficult to define their experience or feels uncomfortable, as suggest by the anxiety and tentative markers from LIWC2015. Cardeña and Pekala (2014) highlights the issue of misinformation due to inadequate metrics, emphasising that if we do not acknowledge missing data, we do not have the full picture. This includes the value and meaning of silences during the mentation process.

One recurring finding that this analysis did not address is the lack of compelling evidence for the concept of an internal attention state. Numerous researchers, including

Carpenter (2004), Dalton (1997), Stanford and Frank (1991), Stanford et al. (1989a, 1989b), found no evidence that measures of absorption or verbal markers designed to assess internal attention reliably predict task outcomes. More recently, similar trends have been observed with other absorption and altered states of consciousness measures. For instance, Marcusson-Clavertz and Cardeña (2011) found no relationship between session *z*-score and attention (which includes the Absorption sub-dimension) via the Phenomenology of Consciousness (PCI; Pekala, 1991), while self-awareness was negatively associated with session outcome. Additionally, Cardeña and Marcusson-Clavertz (2020) found no support for attention via the PCI in the ganzfeld and Watt et al. (2020) found no relationship between any PCI measure and session *z*-score.

These findings suggest that an internal attention state may not be a necessary requirement, nor does it necessarily occur, in the psi ganzfeld. The lack of substantial support from language and various state assessments challenge the underlying theory of the psi ganzfeld paradigm (see Wackermann et al., 2008; Wackermann et al., 2002). As concluded by Vaitl et al. (2005), the ganzfeld is similar to sensory deprivation, but the physical level of sensory stimulation is kept at medium-to-high levels and the ganzfeld imagery is similar to hypnagogic imagery. This is perhaps shown in the LIWC2015 results presented above, and future researchers may wish to examine verbal markers in ganzfeld and non-ganzfeld studies to see if there are any unique, or similar, patterns with the psi ganzfeld task.

Future research should also consider examining changes between the mentation and mentation review periods. Although we did not record the mentation review in KPU study 1074, based on Pooley et al. (2023)'s (Chapter 3) finding that mentation review may decrease study success by adding more noise during judging, the review period process has been identified as a form of retrospective interview about the conscious experience between the participant and the experiment (Wooffitt & Holt, 2011). Analysing this period may reveal how certain characteristics change, likely shifting to past tense and increasing the distance between the experience and experiences, as noted by Wooffitt and Holt (2011), Wooffitt et al. (2010) through "you said" statements. By comparing the mentation and mentation review periods, we may be able to better identify distinct verbal markers which only appear during the ganzfeld period versus those that occur outwith. Further, future research could use the linguistic features of hit and miss sessions as a foundation for hypothesis-driven, confirmatory analyses to better understand the experiences better associated with successful ganzfeld performance.

Overall, this initial exploration using quantitative language analysis to examine ganzfeld mentation reports provides new insights and a baseline for future mentation research. It highlights the complexity and uniqueness of the ganzfeld mentation and underscores the importance of considering all aspects of participants' reports, including silences, to gain a full understanding of their experiences.

7.6 Conclusion

Quantitative language analysis was used to investigate psi ganzfeld mentations collected in two psi ganzfeld precognition studies for the first time. Results show that participants, on average, are ambivalent-to-anxious during the ganzfeld stimulation, are tentative and unsure of their experience but are honest in their descriptions. The high use of first-person pronouns and present tense indicates that participants are actively engaging with the task. The high prevalence of visual and cognitive processing markers confirms earlier assessments of the verbal reports from the ganzfeld experiment. Yet, there are no clear characteristics of participant mentations who scored a hit or a miss. Exploratory linear modelling revealed 5 predictors, none of which have been previously identified in the literature.

This novel analysis adds weight to previous reports of a lack of alteration in the state of consciousness of the participant (Roe, 2009; Wackermann et al., 2002, 2008) and warrants further discussion, especially in a field that does not have a working theory of what is under examination.

Chapter 8

Conclusions

8.1 Thesis summary

This thesis has examined the psi ganzfeld literature as a valuable case study for exploring broader issues in research methodology, particularly in light of the replication crisis and ongoing debate surrounding scientific practice in psychology. As outlined in the introduction, the unique nature of the ganzfeld paradigm – centred on a phenomenon that may not exist – is frequently under debate and has a long history of methodological scrutiny, critical dialogue, and claims of replicable results followed by critiques citing flawed methods or statistical procedures (e.g., Hyman, 2010; Reber & Alcock, 2020; Wagenmakers et al., 2011). Whilst other researchers argue that psi effects have been repeatedly observed under controlled conditions (e.g., Cardeña, 2018, 2025; Storm & Tressoldi, 2020).

Initially, this project intended to conduct a large-scale telepathy ganzfeld experiment incorporating best practices developed in wake of the replication crisis (Kennedy, 2016; Schooler et al., 2018; Wagenmakers et al., 2011). However, due to the COVID-19 pandemic, the focus shifted to a critical assessment of the psi ganzfeld literature itself. This transformed the thesis to empirically investigate and demonstrate the value of the psi research as a broader reflection of the discourse around scientific practices, especially within controversial topic areas. The ganzfeld paradigm serves as a case study and methodological 'ritual' without a clear theoretical framework (Braude, 1992; Rabeyron, 2020; Wiggins & Christopherson, 2019). Although this thesis does not explore whether psi exists, it highlights how the psi literature illustrates persisting tensions within psychological research: between transparency and bias, replication and interpretation, and belief and evidence. The cycle of methodological innovation followed by critique, and eventual return to trusted methods like the psi ganzfeld, demonstrates the rich interplay between theory, practice and belief within this contentious field (Alcock, 2003; Reber & Alcock, 2019, 2020).

8.2 Researcher Degrees of Freedom or Questionable Research Practices?

Both proponents and skeptics, as well as researchers in general, are vulnerable to concerns related to researcher degrees of freedom, questionable research practices and selective reporting. In the case of psi research, these issues are magnified by the absence of a widely accepted theoretical framework and the polarised nature of the field. This lack of consensus often results in the field being labelled as 'deviant' and reinforces binary classifications of researchers as believers or critics (McClenon, 1986; Rabeyron, 2020). As Wiggins and Christopherson (2019) note, methodological reformers have introduced standards to improve the reliability of replication efforts. However, as Brandt et al. (2014) state, every replication inevitably involves some deviation from the original study. One critique of questionable research practices is that researchers frequently fail to provide sufficient detail about the specific methodological decisions they make – decisions which can meaningfully influence study outcomes (John et al., 2012; Wicherts et al., 2016).

This issue is central to this thesis. The psi ganzfeld's rich history of debate, methodological revision and public discourse makes it an ideal case for highlighting and examining degrees of freedom and QRPs. This investigation was conducted through multiple empirical chapters, each examining different aspects of researcher flexibility and transparency.

Chapter 3 presented a meta-regression of previous telepathy ganzfeld studies, analysing how specific study design features influence study outcome. It was the first such analysis in the literature to examine the methodological variables influencing hit rates in telepathy ganzfeld studies. I found that if the sender can hear the receiver, there was an associated 7% increase in hit rate, while the use of a formal mentation review corresponded to a 10% decrease. By conducting a quantitative analysis, we can understand how much these researcher degrees of freedom (flexibility in the study design) directly influence telepathy study outcome. However, the way in which I coded and analysed the meta-regression is also a product of researcher degree of freedom, such as my treatment of outliers and study inclusion decisions (e.g., John et al., 2012; Panagiotou & Ioannidis, 2012; Voracek et al., 2019).

Developing my concerns from conducting a meta-analysis, Chapter 4 explored the ambiguating of how to conduct a meta-analysis, such as inclusion criteria and analytic choices. Given the contentious nature of psi research I wanted to explore published meta-

analyses of the ganzfeld literature and understand how these had been constructed. While previous overviews (Cardeña, 2018, 2025; Tressoldi & Storm, 2021a) have summarised meta-analyses, they rarely scrutinised the methodological heterogeneity between studies. This chapter revealed noticeable variation in which studies were included and how analyses were conducted, known factors that have considerable influence on meta-analysis outcome (Goodyear-Smith et al., 2012; Voracek et al., 2019). Moreover, the average methodological quality of these meta-analyses was found to be low, albeit with the caveat that many predate modern reporting standards such as PRISMA. Regardless, the heterogeneity of meta-analyses is often overlooked and there is increasing concern in the wider research community about how meta-analyses are increasingly produce and seen as the "final word" in academic debates (Taylor & Munafò, 2016). By using the psi ganzfeld literature as an example, we can see how a field that has been conducting replications for nearly 50 years still faces debate about how a meta-analysis *should* or *could* be conducted (Milton, 1999; Schmeidler & Edge, 1999). Again, this chapter questions about when variability in meta-analytic decisions crosses the line into QRPs. The conclusion of the chapter points to promising new methods, such as multiverse type analyses (Plessen et al., 2023; Voracek et al., 2019) and preregistering of meta-analyses to avoid post-hoc analyses of published studies which can be used as an argument of QRPs (Watt & Kennedy, 2017).

Chapter 5 addressed concerns raised in Chapter 2 about experimental software validation (Kennedy, 2016), particularly given historical critiques of psi research citing sensory leakage or flawed randomisation (e.g., Honorton, 1985; Wiseman et al., 1994). As modern psi ganzfeld studies use independently designed experimental software, a common critique is that the software does not truly prevent bias, unintentional or otherwise. Using data from KPU Study 1074, the chapter documented how independent validation confirmed the randomness of the software's number generator and the absence of bias. Due to the persistent debate surrounding psi research, this chapter demonstrates why the psi ganzfeld is a valuable case study to demonstrate the practice of methodological 'belt-tightening' (Wiggins & Christopherson, 2019). However, such validation should ideally occur prior to data collection but this chapter offers a rare example of post-hoc transparency in experimental psychology.

Chapter 6 investigated the impact of analytical flexibility even in preregistered studies, using a specification curve analysis of KPU Study 1074. Despite preregistration, substantial variation remained in how models and variables could be constructed to address a singular hypothesis. The analysis showed that only a small number of models reached

statistical significance, illustrating how researchers could, in principle, select a model structure that aligns with their expectations or belief. Multiverse type analyses can be useful to help process-orientated research, such as the psi ganzfeld and uncover which factors may predict session outcome. But they also demonstrate that a researcher can cherry pick certain modelling structures to present a result that confirms their belief, of any kind.

Chapter 7 focused on underutilised data generated in ganzfeld studies: the verbal mentation reports of participants. The mentation is produced by the participant undergoing the ganzfeld stimulation and is often recorded by researchers (Schmeidler & Edge, 1999). This exploratory linguistic analysis found no distinct verbal features that reliably distinguished successful from unsuccessful trials. Due to lack of a working theory in psi research, the paradigm is regarded as ritual (Rabeyron, 2020) and this verbal analysis further challenges assumptions that the ganzfeld produces an altered state of consciousness (Braude, 1992; Roe, 2009; Wackermann et al., 2002). Participants' language reflected confusion, high levels of analytical thinking, and novelty rather than relaxation and enjoyment. While the implications of this analysis remain uncertain, it introduces a novel approach and highlights how longstanding assumptions in psi research may not hold with novel investigation. It is unclear how future psi research will integrate the language analysis, perhaps to help select future participants or theory development, but I suspect it will become another variable that will be debated considering this chapter found no distinguishable patterns between those who scored a hit or a miss.

8.3 Broader Implications and Future Directions

The issues explored in this thesis, researcher degrees of freedom, questionable research practices, and methodological transparency extend beyond the confines of psi research. In many respects, psi ganzfeld research functions as a barometer for wider psychology research as it regularly attracts interrogation that other fields do not. The adversarial nature of the psi debate encourages extreme scrutiny and reflection.

Researcher degrees of freedom and questionable research practices are routinely discussed in psychological literature (e.g., John et al., 2012; Simmons et al., 2011; Van Elk et al., 2015; Wagenmakers et al., 2021; Wicherts et al., 2016) yet there remains little clarity on how to identify when these practices become problematic. Both psychologists and parapsychologists should continue to prioritise methodological rigour but there remain theoretical concerns, especially in contentious fields like psi research. As highlighted by Flis

(2019) and Wiggins and Christopherson (2019), the replication crisis can be better understood as a reform movement, one that prioritises procedural rigour over theoretical or philosophical considerations. Flis (2019) further argues that psychology demonstrates a 'bottom-up' approach of defining what is 'acceptable' or 'deviant', informed by the researchers' themselves and their own experiences. But this 'indigenous epistemology' is susceptible to human behaviour such as bias and irrationality. Likewise, another concern for psychology research in general is the frequent disconnect between empirical findings and theoretical progress (Feest, 2024; Lavelle, 2022). Without a guiding theoretical framework, even robust empirical efforts will struggle to gain credibility, such as psi research.

And so, future interdisciplinary dialogue between psychologists and philosophers of science may promote better clarity surrounding the epistemological status of practices currently labelled as questionable. By using the psi ganzfeld research as a case-study it highlights the need for philosophical engagement to better understand when degrees of freedom become QRPs, how such boundaries might be drawn and improving the derivation chain between theory and experiment (Scheel et al., 2021).

8.4 Conclusions

The psi ganzfeld paradigm occupies a unique and paradoxical role in science – it is both a staple of replication and a 'deviant science' (McClenon, 1986) conducted by 'academic jesters' (Wagenmakers et al., 2015). As argued by Wicherts et al. (2016), the term 'questionable research practices' may be too severe for most researchers, yet the psi literature frequently attracts this label from parapsychologists (Bierman et al., 2016) and non-parapsychologists alike (Reber & Alcock, 2020; Wagenmakers et al., 2011).

As shown throughout the empirical chapters, variability exists even in direct replications. As demonstrated in Chapter 3, minor methodological differences in telepathy study designs can be regarded as standard researcher degrees of freedom as they are all loosely examining psi in a controlled laboratory setting, trying to produce an altered state of consciousness using a design colloquially agreed to be 'telepathy' (Cardeña, 2018; Rabeyron, 2020). But, this changes to concerns of questionable research practices when others do not report similar findings, as demonstrated in Chapter 4. This perpetuating cycle of calling out too much heterogeneity and QRPs results in continuous methodological and statistical 'belt-tightening' (Wiggins & Christopherson, 2019). Some argue that QRPs are legitimate practices that become problematic only when informed by knowledge of their effect on

outcomes (John et al., 2012; Wiggins & Christopherson, 2019). Yet, given their prevalence in psychology in general and the lack of transparency in reporting, identifying the threshold remains a challenge. The psi ganzfeld literature provides a unique lens through which to examine these evolving norms, illustrating how even replicable findings may not be accepted as true, depending on the theoretical and cultural context (Schooler et al., 2018). In agreement with Wiggins and Christopherson (2019), this thesis argues for greater contextual awareness in interpreting scientific results. Even as practices become more transparent through preregistration, data sharing, and reporting guidelines, the interpretive burden remains with individual researchers. As psychology continues to reflect on its own epistemological foundations, the psi ganzfeld literature offers both a cautionary tale and a valuable framework for engaging with some of its most enduring challenges.

In conclusion, the psi ganzfeld paradigm is more than a controversial research area—it exemplifies the academic discourse about the very nature of psychological science. By using it as a case study, this thesis contributes to ongoing discussions about the standards, ethics, and goals of psychological science. Whether psi exists or not, the lessons drawn from this literature are essential for understanding how science operates, evolves, and sometimes resists change.

References

Alcock, J. E. (2003). Give the null hypothesis a chance: Reasons to remain doubtful about the existence of psi. *Journal of Consciousness Studies*, *10*, 29–50.

Bambra, C., Gibson, M., Sowden, A. J., Wright, K., Whitehead, M., & Petticrew, M. (2009). Working for health? Evidence From Systematic Reviews on the Effects on Health and Health Inequalities of Organisational Changes to the Psychosocial Work Environment. *Preventive Medicine*, *48*(5), 454–461. https://doi.org/10.1016/j.ypmed.2008.12.018

Bancel, P. A. (2018). Simulating Questionable Research Practices. *Proceedings of Presented Papers: The Parapsychological Association 61st Annual Convention*. https://doi.org/10.13140/RG.2.2.12941.64487

Baptista, J., & Derakhshani, M. (2014). Beyond the Coin Toss: Examining Wiseman's Criticisms of Parapsychology. *Journal of Parapsychology*, *78*(1), 56–79.

Baptista, J., Derakhshani, M., & Tressoldi, P. E. (2015). Explicit Anomalous Cognition: A Review of the Best Evidence in Ganzfeld, Forced-Choice, Remote Viewing and Dream Studies. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century* (pp. 192–214). McFarland & Company, Inc.

Bartoš, F., Maier, M., Shanks, D. R., Stanley, T. D., Sladekova, M., & Wagenmakers, E.-J. (2023). Meta-analyses in Psychology Often Overestimate Evidence For and Size of Effects. *Royal Society Open Science*, *10*(7), 230224. https://doi.org/10.1098/rsos.230224

Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bem, D. J., & Honorton, C. (1994). Does Psi Exist? Replicable Evidence for an Anomalous Process of Information Transfer. *Psychological Bulletin*, *115*(1), 4–18. https://doi.org/10.1037/0033-2909.115.1.4

Bem, D. J., Palmer, J., & Broughton, R. S. (2001). Updating the Ganzfeld Database: A Victim of Its Own Success? *Journal of Parapsychology*, *65*(3), 207–218.

Berger, R., & Honorton, C. (1986). An Automated psi Ganzfeld testing system. In D. H. Weiner & D. I. Radin (Eds.), *Research in Parapsychology 1985: Abstracts and Papers from the Twenty-eight Annual Convention of the Parapsychological Association, 1985* (pp. 85–88). The Scarecrow Press, Inc.

Bertini, M., Lewis, H. B., & Witkin, H. A. (1964). Some Preliminary Observations with an Experimental Procedure for the Study on Hypnagogic and Related Phenomena. *Archivio Di Psicologia, Neurologia e Psichiatria*, *25*, 493–534.

Bierman, D. J., Bosga, D. J., Gerding, H., & Wezelman, R. (1993). Anomolous Information Access in the Ganzfeld: Utrecht Novice Series I and II. *Proceedings of the 36th Annual Convention of the Parapsychological Association*, 192–204.

Bierman, D. J., Spottiswoode, J. P., & Bijl, A. (2016). Testing for Questionable Research Practices in a Meta-Analysis: An Example from Experimental Parapsychology. *PloS One*, *11*(5), e0153049. https://doi.org/10.1371/journal.pone.0153049

Bitbol, M., & Petitmengin, C. (2013). A Defense of Introspection from Within. *Constructivist Foundations*, *8*(3), 269–279.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining Psychokinesis: The Interaction of Human Intention with Random Number Generators: A Meta-Analysis. *Psychological Bulletin*, *132*(4), 497–523. https://doi.org/10.1037/0033-2909.132.4.497

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's Threat to Organizational Research: Evidence From Primary and Meta-Analytic Sources. *Personnel Psychology*, *69*(3), 709–750. https://doi.org/10.1111/peps.12111

Boyd, R. L., & Schwartz, H. A. (2021). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Braud, L. W., & Braud, W. G. (1974). Further Studies of Relaxation as a Psi-Conducive State. *Journal of the American Society for Psychical Research*, *68*, 229–245.

Braud, W. G. (2002). Psi-Favourable Conditions. In V. G. Rammohan (Ed.), *New Frontiers of Human Science: A Festschrift for K. Ramakrishna Rao* (pp. 95–118). McFarland & Company, Inc.

Braud, W. G., & Braud, L. W. (1974). Studies of Psi-Facilitating States: Hypnosis, Muscular Relaxation, and an Experimentally Induced Hypnagogic State. *Proceedings of the First International Congress of Pararpsychology and Psychotronics*, *1*, 204–207.

Braude, S. E. (1992). Psi and the Nature of Abilities. *Journal of Parapsychology*, *56*(3), 205–229.

Braude, S. E. (2021). Parra and the Journal of Scientific Exploration. *Journal of Scientific Exploration*, *35*(3), 642–645. https://doi.org/10.31275/20212247

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Broughton, R. S. (1979). Repeatability and experimenter effect: Are subjects really necessary? *Parapsychological Review*, *10*, 11–14.

Broughton, R. S. (1991). *Parapsychology: The controversial science*. Ballantine Books.

Broughton, R. S., & Alexander, C. H. (1997). Autoganzfeld II: An Attempted Replication of the PRL Ganzfeld Research. *Journal of Parapsychology*, *61*(3), 224–226.

Brugger, P., & Taylor, K. I. (2003). ESP Extrasensory Perception or Effect of Subjective Probability? *Journal of Consciousness Studies*, *10*(7), 221–267.

Bulkeley, K., & Graves, M. (2018). Using the LIWC Program to Study Dreams. *Dreaming*, *28*(1), 43–58. https://doi.org/10.1037/drm0000071

Cardeña, E. (2004). Introspection is Alive and Well: Current Methodologies to Study Conscious Experience. *5 Simpósio Da Fundaçâo Bial: Aquém e Alem: Do Cérebro/Behind and Beyond the Brain*, 43–54.

Cardeña, E. (2018). The Experimental Evidence for Parapsychological Phenomena: A Review. *American Psychologist*, *73*(5), 663–677. https://doi.org/10.1037/amp0000236

Cardeña, E. (2020). Editorial: Pieces of the Psi Puzzle and a Recipe for Ganzfeld Success. *Journal of Parapsychology*, *84*(1), 5–7.

Cardeña, E. (2021). Alejandro Parra and Dante's Eighth Circle of Hell. *Journal of Scientific Exploration*, *35*(3), 639–641. https://doi.org/10.31275/20212243

Cardeña, E. (2025). What psi research can – and cannot – say about 'mind beyond the brain'. *International Review of Psychiatry*, 1–5. https://doi.org/10.1080/09540261.2025.2466485

Cardeña, E., Lynn, S. J. E., & Krippner, S. E. (Eds.). (2014). *Varieties of Anomalous Experience: Examining the Scientific Evidence* (2nd ed.). American Psychological Association. doi.org/10.1037/14258-000

Cardeña, E., & Marcusson-Clavertz, D. (2020). Changes in State of Consciousness and Psi in Ganzfeld and Hypnosis Conditions. *Journal of Parapsychology*, *84*(1), 66–84. https://doi.org/10.30891/jopar2020.01.07

Cardeña, E., Marcusson-Clavertz, D., & Palmer, J. (2015). Preface: Reintroducing Parapsychology. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century* (pp. 1–12). McFarland & Company, Inc.

Cardeña, E., & Pekala, R. J. (2014). Researching States of Consciousness and Anomalous Experience. In E. Cardeña, S. J. Lynn, & S. Krippner (Eds.), *Varieties of Anomalous Experience: Examining the Scientific Evidence* (Second, pp. 21–56). American Psychological Association.

Carpenter, J. C. (2004). Implicit Measures of Participant's Experiences in the Ganzfeld: Confirmation of Previous Relationships in a New sample. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention.*, 1–11.

Carver, R. (1978). The Case against Statistical Significance Testing. *Harvard Educational Review*, *48*(3), 378–399.

Center for Open Science. (n.d.). *Preregistration*. Retrieved 31 August 2020, from https://www.cos.io/our-services/prereg

Centre for Reviews and Dissemination. (1995). *The Database of Abstracts of Reviews of Effects (DARE): Quality-assessed reviews*. The University of York.

Clark, C. J., & Tetlock, P. E. (2023). Adversarial Collaboration: The Next Science Reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions* (pp. 905–927). Springer International Publishing. https://doi.org/10.1007/978-3-031-29148-7_32

Cohen, J. (1965). Some Statistical Issues in Psychological Research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp. 95–121). McGraw-Hill.

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., Naveh-Benjamin, M., Barrouillet, P., Camos, V., & Logie, R. H. (2020). How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration. *Perspectives on Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415

Dalton, K. (1997). *The Relationship Between Creativity and Anomalous Cognition in the Ganzfeld* [Doctoral Thesis, University of Edinburgh]. https://era.ed.ac.uk/handle/1842/21184

Dalton, K., Morris, R. L., Delanoy, D. L., Radin, D. I., Taylor, R., & Wiseman, R. (1996). Security Measures In an Automated Ganzfeld System. *Journal of Parapsychology*, *60*(3), 129–148.

Del Giudice, M., & Gangestad, S. W. (2021). A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions.

Advances in Methods and Practices in Psychological Science, *4*(1). https://doi.org/10.1177/2515245920954925

Del Prete, G., & Tressoldi, P. E. (2005). Anomalous Cognition in Hypnagogic State with OBE Induction: An Experimental Study. *The Journal of Parapsychology*, *69*(2), 329–339.

Delanoy, D. L. (1988). An Examination of Subject and Agent Mentation in the Ganzfeld. *European Journal of Parapsychology*, *7*(2–4), 135–168.

Delanoy, D. L., Morris, R. L., & Watt, C. (2004). A Study of Free-Response ESP Performance and Mental Training Techniques. *Journal of the American Society for Psychical Research*, *98*(1–2), 28–67.

Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018). Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts*, *12*(3), 304–316. https://doi.org/10.1037/aca0000137

Dudău, D. P., & Sava, F. A. (2021). Performing Multilingual Analysis With Linguistic Inquiry and Word Count 2015 (LIWC2015). An Equivalence Study of Four Languages. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.570568

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Feest, U. (2024). What is the Replication Crisis a Crisis Of? *Philosophy of Science*, 1–15. https://doi.org/10.1017/psa.2024.2

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322

Fox, J. (2004). An Initial Categorization of the Behavior of Senders During Ganzfeld Trials. *The Journal of the American Society for Psychical Research*, *98*(1–2), 68–92.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for Studying Real-Time Mental Processing Using a Computer Mouse-Tracking Method. *Behavior Research Methods*, *42*(1), 226–241. https://doi.org/10.3758/BRM.42.1.226

Gelman, A., & Loken, E. (2016). The Statistical Crisis in Science. In M. Pitici (Ed.), *The Best Writing on Mathematics 2015* (pp. 305–318). Princeton University Press.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. E. (2017). The Influence of Journal Submission Guidelines on Authors' Reporting of Statistics and Use of Open Research Practices. *PLoS ONE*, *12*(4), 1–15. https://doi.org/10.1371/journal.pone.0175583

Goodyear-Smith, F. A., van Driel, M. L., Arroll, B., & Del Mar, C. (2012). Analysis of Decisions Made in Meta-Analyses of Depression Screening and the Risk of Confirmation Bias: A Case Study. *BMC Medical Research Methodology*, *12*(1), 76. https://doi.org/10.1186/1471-2288-12-76

Goulding, A., & Parker, A. (2001). Finding Psi in the Paranormal: Psychometric Measures Used in Research in Paranormal Beliefs/Experiences and in Research on Psi-Ability. *European Journal of Parapsychology*, *16*, 73–101.

Goulding, A., Westerlund, J., Parker, A., & Wackermann, J. (2004). The First Digital Autoganzfeld Study Using a Real-Time Judging Procedure. *European Journal of Parapsychology*, *19*, 66–97.

Greeley, A. (1987). Mysticism Goes Mainstream. *American Health*, *6*(1), 47–49.

Green, S., Benedetti, J., & Crowley, J. (2003). *Clinical Trials in Oncology*. Chapman & Hall.

Gupta, M., & Agrawal, A. (2012). A Comprehensive Review on Systematic and Meta-Analysis Methods. *International Journal of Pharmacy & Life Sciences*, *3*(2), 1470–1474.

Hacking, I. (1988). Telepathy: Origins of Randomization in Experimental Design. *Isis*, *79*(3), 427–451.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1359–1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.

Holt, N., Delanoy, D. L., & Roe, C. A. (2004). Creativity, Subjective Paranormal Experiences and Altered States of Consciousness. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention.*, 433–436.

Holt, N., Simmonds-Moore, C., Luke, D., & French, C. C. (2012). *Anomalistic Psychology*. Palgrave Macmillan.

Honorton, C. (1972). Reported Frequency of Dream Recall and ESP. *Journal of the American Society for Psychical Research*, *66*(4), 369–374.

Honorton, C. (1977). Psi and Internal Attention States. In B. B. Wolman (Ed.), *Handbook of Parapsychology* (pp. 435–472). McFarland.

Honorton, C. (1985). Meta-Analysis of Psi Ganzfeld Research: A Response to Hyman. *Journal of Parapsychology*, *49*(2), 51–91.

Honorton, C. (1995). Impact of the Sender in Ganzfeld Communication: Meta-Analysis and Power Estimates. *Proceedings of Presented Papers: The Parapsychological Association 38th Annual Convention.*, 132–140.

Honorton, C. (1997). The Ganzfeld Novice: Four -predictors of initial ESP performance. *Journal of Parapsychology*, *61*(2), 143–158.

Honorton, C., Berger, R., Varvoglis, M., Quant, M., Derr, P., Schechter, E., & Ferrari, D. (1990). Psi Communication in the Ganzfeld: Experiments with an Automated Testing System and a Comparison with a Meta-Analysis of Earlier Studies. *Journal of Parapsychology*, *54*(2), 99–139.

Honorton, C., & Ferrari, D. (1989). 'Future telling': A Meta-Analysis of Forced-Choice Precognition Experiments, 1935-1987. *Journal of Parapsychology*, *53*, 281–308.

Honorton, C., Ferrari, D., & Bem, D. J. (1998). Extraversion and ESP Performance: A Meta-Analysis and a New Confirmation. *Journal of Parapsychology*, *62*(3), 255–276.

Honorton, C., & Harper, S. (1974). Psi Mediated Imagery and Ideation in An Experimental Procedure for Regulating Perceptual Input. *Journal of the American Society for Psychical Research*, *68*(2), 156–168.

Honorton, C., & Schechter, E. I. (1986). Ganzfeld Target Retrieval With an Automated Testing System: A Model for Initial Ganzfeld Success. *Proceedings of Presented Papers: The Parapsychological Association 29th Annual Convention.*, 401–414.

Hyman, R. (1985). The Ganzfeld Psi Experiment: A Critical Appraisal. *Journal of Parapsychology*, *49*(1), 3–49.

Hyman, R. (1995). Evaluation of the Program on Anomalous Mental Phenomena. *Journal of Parapsychology*, *50*(4), 351–364.

Hyman, R. (2010). Meta-analysis That Conceals More Than It Reveals: Comment on Storm et al. (2010). *Psychological Bulletin*, *136*(4), 486–490. https://doi.org/10.1037/a0019676

Hyman, R., & Honorton, C. (1986). A Joint Communiqué: The Psi Ganzfeld Controversy. *Journal of Parapsychology*, *50*(4), 351–364.

International Committee of Medical Journal. (2020). *Clinical Trials*. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.1004085

Ioannidis, J. P. A. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Ipsos MORI. (2007). *Survey on beliefs*. https://www.ipsos.com/en-uk/survey-beliefs

JASP Team. (2024). *JASP (Version 0.18.3)[Computer software]* [Computer software]. https://jasp-stats.org/

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, M. (1976). On Publication Policy Regarding Non-Significant Results. *European Journal of Parapsychology*, *1*(2), 1–5.

Kanthamani, H., & Broughton, R. S. (1994). Institute for Parapsychology Ganzfeld-ESP experiments: The Manual Series. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, 182–189.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased Over Time. *PLoS ONE*, *10*(8), 1–12. https://doi.org/10.1371/journal.pone.0132382

Kennedy, J. E. (2013). Can parapsychology move beyond the controversies of retrospective meta-analyses? *Journal of Parapsychology*, *77*(1), 21–35.

Kennedy, J. E. (2014a). Bayesian and Classical Hypothesis Testing: Practical Differences for a Controversial Area of Research. *Journal of Parapsychology*, *78*(2), 170–182.

Kennedy, J. E. (2014b). *Experimenter Misconduct in Parapsychology: Analysis Manipulation and Fraud* (pp. 1–13). https://jeksite.org/psi/misconduct.pdf

Kennedy, J. E. (2015). Beware of Inferential Errors and Low Power with Bayesian Analyses: Power Analysis is Needed for Confirmatory Research. *Journal of Parapsychology*, *79*(1), 53–64.

Kennedy, J. E. (2016). Is the Methodological Revolution in Psychology Over or Just Beginning? *Journal of Parapsychology*, *80*(2), 156–168.

Kennedy, J. E. (2017). Experimenter fraud: What Are Appropriate Methodological Standards? *Journal of Parapsychology*.

Kennedy, J. E., & Watt, C. (2018). *How to Plan Falsifiable Confirmatory Research*.

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

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Koestler Parapsychology Unit. (2018). *Exploratory and Confirmatory Analyses*. https://www.koestler-parapsychology.psy.ed.ac.uk/Documents/explore_confirm.pdf

Kübel, S. L., Fiedler, H., & Wittmann, M. (2021). Red Visual Stimulation in the Ganzfeld Leads to a Relative Overestimation of Duration Compared to Green. *PsyCh Journal*, *10*(1), 5–19. https://doi.org/10.1002/pchj.395

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects. *Nature Human Behaviour*, *4*(6), 659–663. https://doi.org/doi.org/10.1038/s41562-020-0864-3

Lavelle, J. S. (2022). When a Crisis Becomes an Opportunity: The Role of Replications in Making Better Theories. *The British Journal for the Philosophy of Science*, *73*(4), 965–986. https://doi.org/10.1086/714812

Lindsay, D. S., & Nosek, B. A. (2018, February 28). Preregistration Becoming the Norm in Psychological Science. *APS Observer*, *31*. https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Marcusson-Clavertz, D., & Cardeña, E. (2011). Hypnotizability, Alterations in Consciousness, and Other Variables as Predictors of Performance in a Ganzfeld Psi Task. *Journal of Parapsychology*, *75*(2), 235–259.

Masur, P. K., & Scharkow, M. (2020). *specr: Conducting and Visualizing Specification Curve Analyses (Version 1.0.1)*. https://CRAN.R-project.org/package=specr

Mauskopf, S. H., & McVaugh, M. R. (1980). *The Elusive Science: Origins of Experimental Psychical Research*. John Hopkins University Press.

May, E. C. (2007). Advances in Anomalous Cognition Analysis: A Judge-Free and Accurate Confidence-Calling. *Proceedings of Presented Papers: The Parapsychological Assocation 50th Annual Convention*, 57–63.

May, E. C., Utts, J. M., & Spottiswoode, J. P. (1995). Decision Augmentation Theory: Toward a Model of Anomalous Mental Phenomena. *Journal of Parapsychology*, *59*(3), 195–220.

McClenon, J. (1986). Scientific Rhetoric and the Ganzfeld Debate. *The Journal of Parapsychology*, *50*(4), 371–375.

Milton, J. (1986). *Displacement Effects, Role of the Agent, and Mentation Categories in Relation to ESP Performance* [Doctoral Thesis, University of Edinburgh]. https://era.ed.ac.uk/handle/1842/6953

Milton, J. (1997a). A Meta-analytic comparison of the Sensitivity of Direct Hits and Sums of Ranks as Outcome Measures for Free-Response Studies. *Journal of Parapsychology*, *61*(3), 277–241.

Milton, J. (1997b). Meta-analysis of Free-Response ESP Studies Without Altered States of Consciousness. *Journal of Parapsychology*, *61*(4), 279–319.

Milton, J. (1999). Should Ganzfeld Research Continue to be Crucial in the Search for a Replicable Psi Effect? Part I. Discussion Paper and Introduction to an Electronic-mail Discussion. *Journal of Parapsychology*, *63*(4), 309–333.

Milton, J., & Wiseman, R. (1997). *Guidelines for Extrasensory Perception Research*. University of Hertfordshire Press.

Milton, J., & Wiseman, R. (1999). Does psi exist? Lack of Replication of an Anomalous Process of Information Transfer. *Psychological Bulletin*, *125*(4), 387–391. https://doi.org/doi.org/10.1037/0033-2909.125.4.387

Moore, D. W. (2005). *Three in Four Americans Believe in Paranormal*. Gallup. https://news.gallup.com/poll/16915/three-four-americans-believe-paranormal.aspx

Morris, R. L., Cunningham, S., McAlpine, S., & Taylor, R. (1993). Toward Replication and Extension of Autoganzfeld results. *Proceedings of the 36th Annual Convention of the Parapsychological Association*, 177–191.

Morris, R. L., Dalton, K., Delanoy, D., & Watt, C. (1995). Comparison of the Sender/No Sender Condition in the Ganzfeld. *Proceedings of the 38th Annual Parapsychological Association Convention*, 244–258.

Morris, R. L., Summers, J., & Yim, S. (2003). Evidence of Anomalous Information Transfer with a Creative Population in Ganzfeld Stimulation. *Proceedings of Presented Papers: The Parapsychological Association 46th Annual Convention*, 116–131.

Muncer, S., Taylor, S., & Craigie, M. (2002). Power Dressing and Meta-Analysis: Incorporating Power Analysis into Meta-Analysis. *Journal of Advanced Nursing*, *38*(3), 274–280.

Murray, A. L. (2011). The Validity of the Meta-Analytic Method in Addressing the Issue of Psi Replicability. *Journal of Parapsychology*, *75*(2), 251–277.

Nahm, M. (2021). A New Case of Scientific Dishonest in the Field of Parapsychology. *Journal of Scientific Exploration*, *35*(3), 623–638.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Oliveira, D., Rosenthal, M., Morin, N., Yeh, K.-C., Cappos, J., & Zhuang, Y. (2014). It's the Psychology Stupid: How Heuristics Explain Software Vulnerabilities and How Priming Can Illuminate Developer's Blind Spots. *Proceedings of the 30th Annual Computer Security Applications Conference*, 296–305. https://doi.org/10.1145/2664243.2664254

Oliveras, I., Losilla, J.-M., & Vives, J. (2017). Methodological Quality is Underrated in Systematic Reviews and Meta-Analyses in Health Psychology. *Journal of Clinical Epidemiology*, *86*, 59–70. https://doi.org/10.1016/j.jclinepi.2017.05.002

Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science*, *349*(6251), 943–951. https://doi.org/10.1126/science.aac4716

Palmer, J. (1992). From Survival to Transcendence: Reflections on Psi as Anomalous. *Journal of Parapsychology*, *56*(3), 229–255.

Palmer, J. (1997). Correlates of ESP Magnitude and Direction in the FRNM Manual Ganzfeld Database. *The Parapsychological Association 40th Annual Convention*, 108–123.

Palmer, J. (2015). Experimental Methods in Anomalous Cognition and Anomalous Perturbation Research. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century* (pp. 49–62). McFarland & Company, Inc.

Palmer, J. (2016). Hansel's ghost: Resurrection of the Experimenter Fraud Hypothesis in Parapsychology. *Journal of Parapsychology*, *80*(1), 5–16.

Palmer, J., & Broughton, R. S. (2000). An Updated Meta-Analysis of Post-PRL ESP Ganzfeld Experiments: The Effect of Standardness. *Journal of Parapsychology*, *64*(3), 249–250.

Palmer, J., Khamashta, K., & Isrealson, K. (1979). An ESP Ganzfeld Experiment with Transcendental Meditators. *Journal of the American Society for Psychical Research*, *73*(4), 333–348.

Panagiotou, O. A., & Ioannidis, J. P. A. (2012). Primary Study Authors of Significant Studies are More Likely to Believe That a Strong Association Exists in a Heterogeneous Meta-Analysis Compared with Methodologists. *Journal of Clinical Epidemiology*, *65*(7), 740–747. https://doi.org/10.1016/j.jclinepi.2012.01.008

Parapsychological Association. (2015, November 17). *Extrasensory Perception (ESP)*. https://www.parapsych.org/articles/53/301/extrasensory_perception_esp.aspx

Parker, A. (2006). A Ganzfeld Study with Identical Twins. *Proceedings of Presented Papers: The Parapsychological Association 49th Annual Convention.*, 330–334.

Parker, A. (2010). A ganzfeld study using identical twins. *Journal of the Society for Psychical Research*, *74.2*(899), 118–126.

Parker, A., Frederiksen, A., & Johansson, H. (1997). Towards specifying the recipe for success with the Ganzfeld. *European Journal of Parapsychology*, *13*, 15–27.

Parker, A., Grams, D., & Pettersson, C. (1998). Further Variables Relating to Psi in the Ganzfeld. *The Journal of Parapsychology*, *62*(4), 319–337.

Parker, A., Persson, A., & Haller, A. (2000). Using Qualitative Reserach for Theory Development: Top Down Processes in Psi-Mediation. *Journal of the Society for Psychical Research*, *64*, 65–81.

Parker, A., & Westerlund, J. (1998). Current Research in Giving the Ganzfeld an Old and a New Twist. *Proceedings of Presented Papers: The Parapsychological Association 41st Annual Convention*, 135–142.

Parra, A., & Argibay, J. C. (2007). "Token-object" Effect and Medical Diagnosis: An Experimental Study. *Proceedings of Presented Papers: The Parapsychological Association 50th Annual Convention*, 95–102.

Pechey, R., & Halligan, P. (2011). The Prevalence of Delusion-like Beliefs Relative to Sociocultural Beliefs in the General Population. *Psychopathology*, *44*(2), 106–115. https://doi.org/10.1159/000319788

Pechey, R., & Halligan, P. (2012). Prevalence and Correlates of Anomalous Experiences in a Large Non-Clinical Sample. *Psychology and Psychotherapy*, *85*(2), 150–162. https://doi.org/10.1111/j.2044-8341.2011.02024.x

Pehlivanova, M., Weiler, M., & Greyson, B. (2024). Cognitive Styles and Psi: Psi Researchers are More Similar to Skeptics Than to Lay Believers. *Frontiers in Psychology*, *15*, 1398121. https://doi.org/10.3389/fpsyg.2024.1398121

Pekala, R. J. (1991). *Quantifying Consciousness*. Springer US. https://doi.org/10.1007/978-1-4899-0629-8

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015* [Operator's Manual]. Pennebaker Conglomerates.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin.

Petitmengin, C., & Bitbol, M. (2009). The Validity of First-Person Descriptions as Authenticity and Coherence. *Journal of Consciousness Studies*, *16*(10–12), 252–284.

Plessen, C. Y., Karyotaki, E., & Cuijpers, P. (2022). Exploring the Efficacy of Psychological Treatments for Depression: A Multiverse Meta-Analysis Protocol. *BMJ Open*, *12*(e050197). https://doi.org/10.1136/bmjopen-2021-050197

Plessen, C. Y., Karyotaki, E., Miguel, C., Ciharova, M., & Cuijpers, P. (2023). Exploring the efficacy of psychotherapies for depression: A multiverse meta-analysis. *BMJ Mental Health*, *26*(1), e300626. https://doi.org/10.1136/bmjment-2022-300626

Polanin, J. R., Hennessy, E. A., & Sho, T. (2020). Transparency and Reproducibility of Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

Pooley, A. L. (2021). Sender-Receiver Relationship in the Ganzfeld. In *Book of Abstracts: 44th International Annual Conference of the Society for Psychical Research*. Society for Psychical Research.

Pooley, A. L., Murray, A. L., & Watt, C. (2023). Understanding the Factors at Play in the Sender-Receiver Dynamic During the Telepathy Ganzfeld: A Meta-Analysis. *Journal*

of Anomalous Experience and Cognition, *3*(1), 42–77. https://doi.org/10.31156/jaex.23878

Rabeyron, T. (2020). Why Most Research Findings About Psi Are False: The Replicability Crisis, the Psi Paradox and the Myth of Sisyphus. *Frontiers in Psychology*, *11*(11:562992).

Radin, D. (2006). Conscious Psi. In *Entangled Minds: Extrasensory experiences in a quantum reality* (pp. 98–130). Paraview Pocket Books.

Rauvola, R. S., & Rudolph, C. W. (2023). Worker Aging, Control, and Well-being: A Specification Curve Analysis. *Acta Psychologica*, *233*, 103833. https://doi.org/10.1016/j.actpsy.2023.103833

Reber, A. S., & Alcock, J. E. (2019). Why Parapsychological Claims Cannot Be True. *Skeptical Inquirer*, *43*(4), 8–10.

Reber, A. S., & Alcock, J. E. (2020). Searching for the Impossible: Parapsychology's Elusive Quest. *American Psychologist*, *75*(3), 391–399. https://doi.org/10.1037/amp0000486

Rhine, J. B. (1974). A New Case of Experimenter Unreliability. *Journal of Parapsychology*, *38*(1), 215–225.

Rhine, J. B. (1975). Second Report on a Case of Experimenter Fraud. *Journal of Parapsychology*, *39*, 306–325.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Roe, C. A. (2009). The Role of Altered States of Consciousness in Extrasensory Experiences. In *Anomalous Experiences: Essays from Parapsychological and Psychological Perspectives*. McFarland & Company, Inc.

Roe, C. A., & Holt, N. (2005). A further consideration of the sender as a PK agent in ganzfeld ESP studies. *Journal of Parapsychology*, *69*(1), 113–127.

Roe, C. A., Holt, N., & Simmonds, C. A. (2003). Considering the sender as a PK agent in the Ganzfeld ESP studies. *Journal of Parapsychology*, *67*(1), 129–145.

Roe, C. A., McKenzie, E. A., & Ali, A. N. (2001). Sender and Receiver Creativity Scores as Predictors of Performance at a Ganzfeld ESP Task. *Journal of the Society for Psychical Research*, *65*(2), 107–121.

Roe, C. A., Sherwood, S. J., & Holt, N. (2004). Interpersonal Psi: Exploring the Role of the Sender in Ganzfeld Gesp Tasks. *Journal of Parapsychology*, *68*(2), 361–380.

Roney-Dougal, S. (2015). Ariadne's Thread: Meditation and Psi. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century* (pp. 125–138).

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes Factor Meta-Analysis of Recent Extrasensory Perception Experiments: Comment on Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*(1), 241–247. https://doi.org/10.1037/a0029008

Runco, M. A., Plucker, J. A., & Lim, W. (2001). Development and Psychometric Integrity of a Measure of Ideational Behavior. *Creativity Research Journal*, *13*(3–4), 393–400. https://doi.org/10.1207/S15326934CRJ1334_16

Sargent, C. L. (1980). *Exploring Psi in the Ganzfeld* (17; Parapsychological Monographs). Parapsychology Foundation.

Schalken, N., & Rietbergen, C. (2017). The Reporting Quality of Systematic Reviews and Meta-Analyses in Industrial and Organizational Psychology: A Systematic Review. *Frontiers in Psychology*, *8*, 1395. https://doi.org/10.3389/fpsyg.2017.01395

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schlitz, M., & Honorton, C. (1992). Ganzfeld Psi Performance within an Artically Gifted Population. *The Journal of the American Society for Psychical Research*, *86*(2), 83–98.

Schlitz, M., & Radin, D. (2003). Non-sensory access to information: The ganzfeld studies. In *Healing, Intention and Energy Medicine* (pp. 75–82). Elsevier. https://doi.org/10.1016/B978-0-443-07237-6.50012-7

Schmeidler, G. R., & Edge, H. (1999). Should Ganzfeld Research Continue to be Crucial in the Search for a Replicable Psi Effect? Part II. Edited Ganzfeld Debate. *Journal of Parapsychology*, *63*(4), 355–388.

Schmidt, S., Schneider, R., Utts, J., & Walach, H. (2004). Distant Intentionality and the Feeling of Being Started At. *British Journal of Psychology*, *95*(2), 235–247. https://doi.org/10.1348/000712604773952449

Schmidt, T. T., & Prein, J. C. (2019). The Ganzfeld Experience—A Stably Inducible Altered State of Consciousness: Effects of Different Auditory Homogenizations. *PsyCh Journal*, *8*(1), 66–81. https://doi.org/10.1002/pchj.262

Schooler, J. W., Baumgart, S., & Franklin, M. (2018). Entertaining Without Endorsing: The Case for the Scientific Investigation of Anomalous Cognition. *Psychology of Consciousness: Theory, Research, and Practice*, *5*(1), 63–77. https://doi.org/10.1037/cns0000151

Schwab, A., & Starbuck, W. H. (2017). A Call for Openness in Research Reporting: How to Turn Covert Practices Into Helpful Tools. *Academy of Management Learning & Education*, *16*(1), 125–141. https://doi.org/10.5465/amle.2016.0039

Sharpe, D. (1997). Of Apples and Oranges, File Drawers and Garbage: Why Validity Issues in Meta-Analysis Will Not Go Away. *Clinical Psychology Review*, *17*(8), 881–901. https://doi.org/10.1016/S0272-7358(97)00056-1

Sharpe, D., & Poets, S. (2020). Meta-Analysis as a Reponse to the Replication Crisis. *Canadian Psychology / Psychologie Canadienne*, *61*(4), 377–387. https://doi.org/10.1037/cap0000215

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and How. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification Curve Analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Sladekova, M., Webb, L. E. A., & Field, A. P. (2023). Estimating the Change in Meta-Analytic Effect Size Estimates After the Application of Publication Bias Adjustment Methods. *Psychological Methods*, *28*(3), 664–686. https://doi.org/10.1037/met0000470

Smith, M. D., & Savva, L. (2004). Experimenter effects and psi performance using a digital autoganzfeld system. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention.*, 461–463.

Smith, S. W., & Greer, B. D. (2022). Validating Human-Operant Software: A Case Example. *Behavior Analysis: Research and Practice*, *22*(4), 389–403. https://doi.org/10.1037/bar0000244

Sochat, V. V., Eisenberg, I. W., Enkavi, A. Z., Li, J., Bissett, P. G., & Poldrack, R. A. (2016). The Experiment Factory: Standardizing Behavioral Experiments. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00610

Spencer, B. (1995). Correlations, Sample Size, and Practical Significance: A Comparison of Psychologyical and Medical Investigations. *The Journal of Psychology*, *129*, 469–475. https://doi.org/10.1080/00223980.1995.9914982

Srinivas, T. R., Ho, B., Kang, J., & Kaplan, B. (2015). Post Hoc Analyses: After the Facts. *Transplantation*, *99*(1), 17–20. https://doi.org/10.1097/TP.0000000000000581

Stanford, R. G., & Frank, S. (1991). Prediction of Ganzfeld ESP-Task Performance from Session-Based Verbal Indicators of Psychological Function: A Second Study. *Journal of Parapsychology*, *55*(4), 349–376.

Stanford, R. G., Frank, S., Kass, G., & Skoll, S. (1989a). Ganzfeld as an ESP-Favorable Setting. Part II: Prediction of ESP-Task Performance Through Verbal-Transcript Measures of Spontaneity, Suboptimal Arousal, and Internal Attention State. *Journal of Parapsychology*, *53*(2), 95–124.

Stanford, R. G., Frank, S., Kass, G., & Skoll, S. (1989b). Ganzfeld as an ESP-Favourable Setting: Part I. Assessment of Spontaneity, Arousal, and Internal Attention State Through Verbal Transcript Analysis. *Journal of Parapsychology*, *53*(1), 1–42.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Storm, L., & Ertel, S. (2001). Does Psi Exist? Comments on Milton and Wiseman's (1999) Meta-Analysis of Ganzfeld Research. *Psychological Bulletin*, *127*(3), 424–433. https://doi.org/10.1037/0033-2909.127.3.424

Storm, L., & Tressoldi, P. E. (2020). Meta-Analysis of Free-Response Studies 2009-2018: Assessing the Noise-Reduction Model Ten Years On. *Journal of the Society for Psychical Research*, *84*(4), 193–219. https://doi.org/10.31234/osf.io/3d7at

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010a). Meta-Analysis of ESP Studies, 1987-2010: Assessing the Success of Forced-Choice Design in Parapsychology. *Journal of Parapsychology*, *76*(2), 243–273.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010b). Meta-Analysis of Free-Response Studies, 1992-2008: Assessing the Noise Reduction Model in Parapsychology. *Psychological Bulletin*, *136*(4), 471–485. https://doi.org/10.1037/a0019457

Storm, L., Tressoldi, P. E., & Utts, J. (2013). Testing the Storm et al. (2010) Meta-Analysis Using Bayesian and Frequentist Approaches: Reply to Rouder et al.(2013). *Psychological Bulletin*, *139*(1), 248–254. https://doi.org/10.1037/a0029506

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688.

Symmons, C., & Morris, R. L. (1997). Drumming at Seven Hz and Automated Ganzfeld Performance. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention.*, 441–453.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., Rooij, I. van, Zandt, T. V., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Targ, E., Schlitz, M., & Irwin, H. (2000). Psi-related experiences. In E. Cardeña, S. J. Lynn, & S. Krippner (Eds.), *Varieties of Anomalous Experience: Examining the Scientific Evidence* (pp. 219–252). American Psychological Association.

Tart, C. T. (1978). Psi functioning and altered states of consciousness: A perspective. In B. Shapin & L. Coly (Eds.), *Psi and states of awareness* (pp. 180–210). Parapsychology Foundation.

Taylor, A. E., & Munafò, M. R. (2016). Triangulating Meta-Analysis: The Example of the Serotonin Transporter Gene, Stressful Life Events and Major Depression. *BMC Psychology*, *4*(1), 23. https://doi.org/10.1186/s40359-016-0129-0

Thalbourne, M. A. (2010). The Australian Sheep-goat Scale: Development and Empirical Findings. *Australian Journal of Parapsychology*, *10*(1), 5–39. https://doi.org/10.3316/informit.215454028272587

Tressoldi, P. E. (2016). Prospective Statistical Power: Sample Size Recommendations for the Investigation of the Main Parapsychological Phenomena. *Journal of Scientific Exploration*, *30*(1), 10–15.

Tressoldi, P. E. (2019). *Ganzfeld database 1974-2018*. https://open-data.spr.ac.uk/dataset/1974-2018-ganzfeld-database

Tressoldi, P. E., & Del Prete, G. (2007). ESP Under Hypnosis: The Role of Induction Instructions and Personality Characteristics. *The Journal of Parapsychology*, *71*(1), 125–137.

Tressoldi, P. E., & Storm, L. (2021a). Anomalous Cognition: An Umbrella Review of the Meta-Analytic Evidence. *Journal of Anomalous Experience and Cognition*, *1*(1–2), 55–72. https://doi.org/10.31156/jaex.23206

Tressoldi, P. E., & Storm, L. (2021b). Stage 1 Registered Report: Anomalous perception in a Ganzfeld condition—A meta-analysis of more than 40 years investigation [version 3]. *F1000Research*, *9*(826). https://doi.org/10.12688/f1000research.24868.3

Tressoldi, P. E., & Storm, L. (2023). Stage 2 Registered Report: Anomalous Perception in a Ganzfeld Condition—A Meta-Analysis of More Than 40 Years Investigation. *F1000Research*, *10*, 234. https://doi.org/10.12688/f1000research.51746.2

Tressoldi, P. E., & Utts, J. (2015). Statistical Guidelines for Empirical Studies. In E. Cardeña, J. Palmer, & D. Marcusson-Clavertz (Eds.), *Parapsychology: A Handbook for the 21st Century* (pp. 83–93). McFarland & Company, Inc.

Ullman, M., Krippner, S., & Vaughan, A. (2002). *Dream Telepathy: Experiments in Nocturnal ESP*. Hampton Roads.

U.S Food and Drug Administration. (1998). *E9 Statistical Principles for Clinical Trials*. https://www.fda.gov/media/71336/download

U.S Food and Drug Administration. (2002). *General Principles of Software Validation*.

Utts, J. (1991). Replication and Meta-analysis in Parapsychology. *Statistical Science*, *6*(4), 363–403.

Utts, J., Norris, M., Suess, E., & Johnson, W. (2010). The Strength of Evidence Versus the Power of Belief: Are we all Bayesians? In C. Reading (Ed.), *Data and Context in Statistics education: Towards an evidence-based society. Proceedings of the Eight International Conference on Teaching Statistics.* https://doi.org/10.1198/000313007X192563

Vaitl, D., Gruzelier, J., Jamieson, G. A., Lehmann, D., Ott, U., Sammer, G., Strehl, U., Birbaumer, N., Kotchoubey, B., Kübler, A., Miltner, W. H. R., Pütz, P., Strauch, I., Wackermann, J., & Weiss, T. (2005). Psychobiology of Altered States of Consciousness. *Psychological Bulletin*, *131*(1), 98–127. https://doi.org/10.1037/0033-2909.131.1.98

Van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-Analyses Are no Substitute for Registered Replications: A Skeptical

Perspective on Religous Priming. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01365

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Viera, A. J., & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, *37*(5), 360–363.

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which Data to Meta-Analyze, and How?: A Specification-Curve and Multiverse-Analysis Approach to Meta-Analysis. *Zeitschrift Fur Psychologie / Journal of Psychology*, *227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Wackermann, J., Pütz, P., & Allefeld, C. (2008). Ganzfeld-induced hallucinatory experience, its phenomenology and cerebral electrophysiology. *Cortex*, *44*(10), 1364–1378. https://doi.org/10.1016/j.cortex.2007.05.003

Wackermann, J., Pütz, P., Büchi, S., Strauch, I., & Lehmann, D. (2002). Brain Electrical Activity and Subjective Experience During Altered States of Consciousness: Ganzfeld and Hypnagogic States. *International Journal of Psychophysiology*, *46*(2), 123–146. https://doi.org/10.1016/S0167-8760(02)00070-3

Wagenmakers, E.-J., & Grünwald, P. (2006). Commentary: A Bayesian Perspective on Hypothesis Testing. A Comment on Killeen (2005). *Psychological Science*, *17*(7), 641–642. https://doi.org/10.1111/j.1467-9280.2006.01757.x

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F.,

Tendeiro, J., & Aczel, B. (2021). Seven steps Toward More Transparency in Statistical Practice. *Nature Human Behaviour*, *5*(11), 1473–1480. https://doi.org/10.1038/s41562-021-01211-8

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R. A., & van der Maas, H. L. J. (2015). A Skeptical Eye on Psi. In E. C. May & S. B. Marwaha (Eds.), *Extrasensory Perception: Support, Skepticism, and Science* (Vol. 1, pp. 153–176). Bloomsbury Publishing USA.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L. J., Yantis, C., Wong, K. M., Sun, J., Prims, J. P., Mueller, A. B., Melton, Z. J., & Carsel, T. S. (2018). Why Do Some Psychology Researchers Resist Adopting Proposed Reforms to Research Practices? A Description of Researchers' Rationales. *Advancies in Methods and Practices in Psychological Sciences*, *1*(2), 166–173. https://doi.org/10.1177/2515245918757427

Watt, C. (2005). Parapsychology's Contribution to Psychology: A View From the Front Line. *Journal of Parapsychology*, *69*(2), 215–231.

Watt, C. (2006). Research Assistants or Budding Scientists? A Review of 96 undergraduate student projects at the Koestler Parapsychology Unit. *Journal of Parapsychology*, *70*(2), 335–356.

Watt, C., & Brady, C. (2002). Experimenter Effects and the Remote Facilitation of Attention Focusing: Two Studies and the Discovery of an Artifact. *The Journal of Parapsychology*, *66*(1), 49–71.

Watt, C., Dawson, E., Tullo, A., Pooley, A., & Rice, H. (2020). Testing Precognition and an Altered State of Consciousness with Selected Participants in the Ganzfeld. *Journal of Parapsychology*, *84*(1), 21–37. https://doi.org/10.30891/jopar.2020.01.05

Watt, C., & Kennedy, J. E. (2015). Lessons From the First Two Years of Operating a Study Registry. *Frontiers in Psychology*, *6*, 1–4. https://doi.org/10.3389/fpsyg.2015.00173

Watt, C., & Kennedy, J. E. (2016). Stimulating Progress in Parapsychology: Prospective Meta-Analysis. *The Parapsychological Association 59th Annual Convention*, 49–60.

Watt, C., & Kennedy, J. E. (2017). Options for Prospective Meta-Analysis and Introduction of Registration-Based Prospective Meta-Analysis. *Frontiers in Psychology*, *7*(2030). https://doi.org/doi: 10.3389/fpsyg.2016.02030

Watt, C., & Tierney, I. (2014). Psi-Related Experiences. In E. Cardeña, S. J. Lynn, & S. Krippner (Eds.), *Varieties of Anomalous Experience: Examining the Scientific Evidence* (pp. 241–272). American Psychological Association.

Westerlund, J., Parker, A., Dalkvist, J., & Hadlaczky, G. (2006). Remarkable Correspondences Between Ganzfeld Mentation and Target Content—A Psychical or Psychological Effect? *Journal of Parapsychology*, *70*(1), 23–48.

Wezelman, R., & Bierman, D. J. (1997). Process Orientated Ganzfeld Research in Amsterdam: Series IV B: Emotionality of Target material, Series V and VI: Judging Procedure and States of Consciousness. *Proceedings of the 40th Annual Parapsychology Association Convention*, 477–492.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*(1832). https://doi.org/doi.org/10.3389/fpsyg.2016.01832

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Williams, B. J. (2011). Revisiting the Ganzfeld ESP Debate: A Basic Review and Assessment. *Journal of Scientific Exploration*, *25*(4), 639–661.

Willin, M. J. (1996a). A Ganzfeld Experiment Using Musical Targets. *Journal of the Society for Psychical Research*, *61*(842), 1–17.

Willin, M. J. (1996b). A Ganzfeld Experiment Using Musical Targets with Previous High Scorers From the General Population. *Journal of the Society for Psychical Research*, *61*(843), 103–106.

Wilson, D. B., & Shadish, W. R. (2006). On Blowing Trumpets to the Tulips: To Prove or Not to Prove the Null Hypothesis—Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin*, *132*(4), 524–528. https://doi.org/10.1037/0033-2909.132.4.524

Wiseman, R. (2010). 'Heads I Win, Tails You Lose': How Parapsychologists Nullify Null Results. *Skeptical Inquirer*, *34*(1), 36–39.

Wiseman, R., Smith, M., & Kornbrot, D. (1994). Assessing Possible Sender-to-Experimenter Acoustic Leakage in the PRL Autoganzfeld. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention.*, 439–454.

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered Reports: An Early Example and Analysis. *PeerJ*, *7*, e6232. https://doi.org/doi.org/10.7717/peerj.6232

Wooffitt, R. (2003). Conversation Analysis and Parapsychology: Experimenter-Subject Interaction in Ganzfeld Experiments. *Journal of Parapsychology*, *67*(2), 299–323.

Wooffitt, R., & Holt, N. (2011). *Looking In and Speaking Out: Introspection, Consciousness, Communication*. Imprint Academic.

Wooffitt, R., Holt, N., & Allistone, S. (2010). Introspection as Institutional Practice: Reflections on the Attempt to Capture Conscious Experience in a Parapsychology Experiment. *Qualitative Research in Psychology*, *7*(1), 5–20. https://doi.org/10.1080/14780880903304568

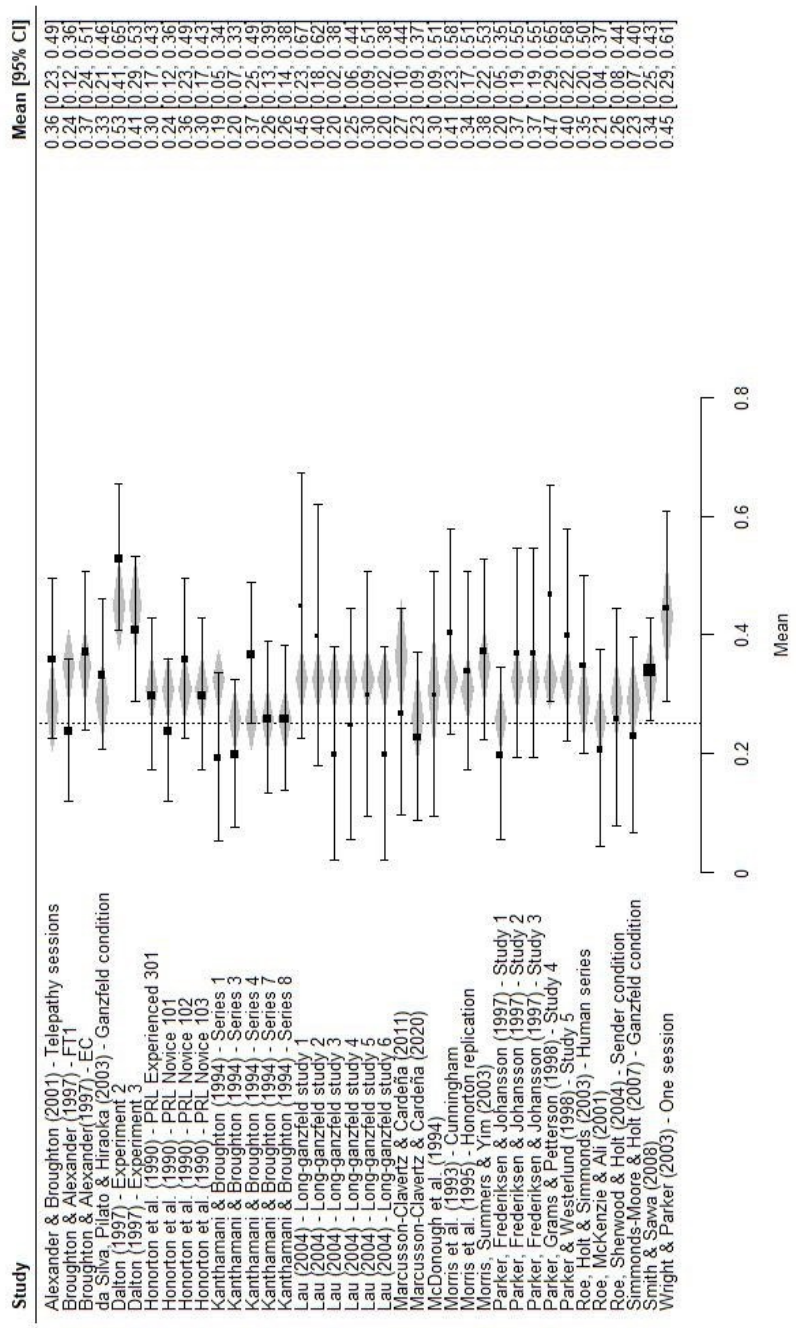
Yuan, K.-H., & Maxwell, S. (2005). On the Post-Hoc Power in Testing Mean Differences. *Journal of Education and Behavioural Statistics*, *30*(2), 141–167.

Appendix A

Chapter 3 Supplemental Materials

Supplemental Figure 1

Forest plot for Model 1.1



Note. Reference line is set to MCE (0.25).

Supplemental Table 1

Model 2: Proportion of hits summary output with the same 3 outliers removed in Model 1.1

	Estimate	Standard error	t-value	p-value	95% CI Lower Bound 95	95% CI Upper Bound
Intercept	.36	.05	7.15	<.0001***	0.26	0.46
See	.00	.03	0.03	.92	-0.07	0.07
Hear	.07	.03	0.03	.03*	0.01	0.13
Hear judging	-.04	.03	0.03	.28	-0.11	0.03
Silent	.02	.03	0.03	.53	-0.04	0.08
Review	-.10	.04	0.04	.02*	-0.19	-0.01

Note. *** indicates significance at the 1% level. * indicates significance at the 5% level,

Supplemental Table 2

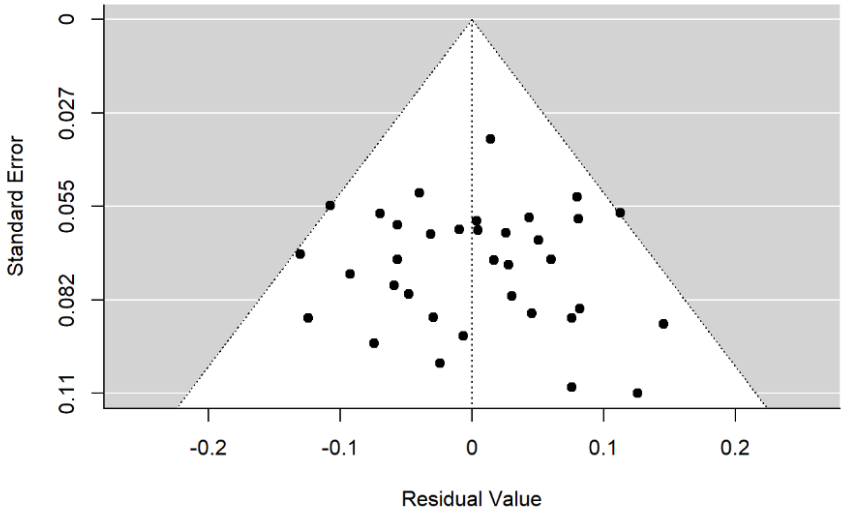
Model 2: Permutation test results (5000 iterations)

	Estimate	Standard error	t-value	p-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.36	.05	7.15	.05*	0.26	0.46
See	.00	.03	0.10	.92	-0.07	0.07
Hear	.07	.03	2.27	.03*	0.01	0.13
Hear judging	-.04	.03	-1.10	.29	-0.11	0.03
Silent	.02	.03	0.64	.53	-0.04	0.08
Review	-.10	.04	-2.38	.03*	-0.19	-0.01

Note. * indicates significance at the 5% level.

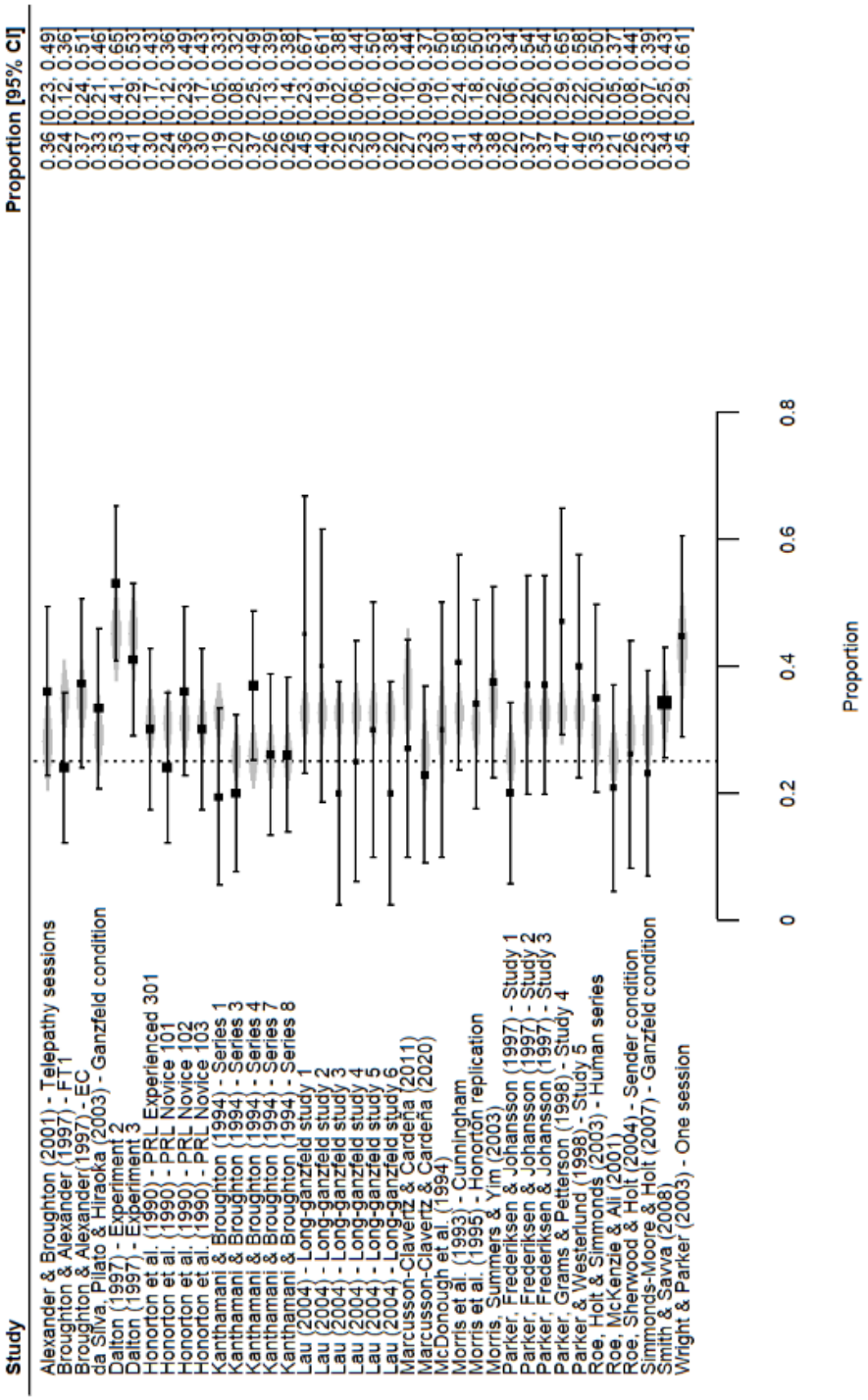
Supplemental Figure 2

Forest plot for Model 2 with same 3 outliers removed from Model 1.1



Supplemental Figure 3

Funnel plot for Model 2



Note. Reference line is set to MCE (0.25).

Supplemental Table 3

Model 1.2: Binomial mean model with the Review factor removed[10]

	Estimate	Standard error	t-value	p-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.23	.02	10.22	<.0001***	0.19	0.28
See	.04	.03	1.51	.14	−0.01	0.10
Hear	.09	.03	3.38	< .01*	0.04	0.15
Hear judging	-.08	.03	−2.70	.01*	−0.14	−0.02
Silent	.03	.03	0.97	.34	-0.03	0.08

Note. *** indicates significance at the 0.1% level. ** indicates significance at the 1% level. * indicates significance at the 5% level.

Supplemental Table 4

Model 1.2: Permutation test (5000 iterations)

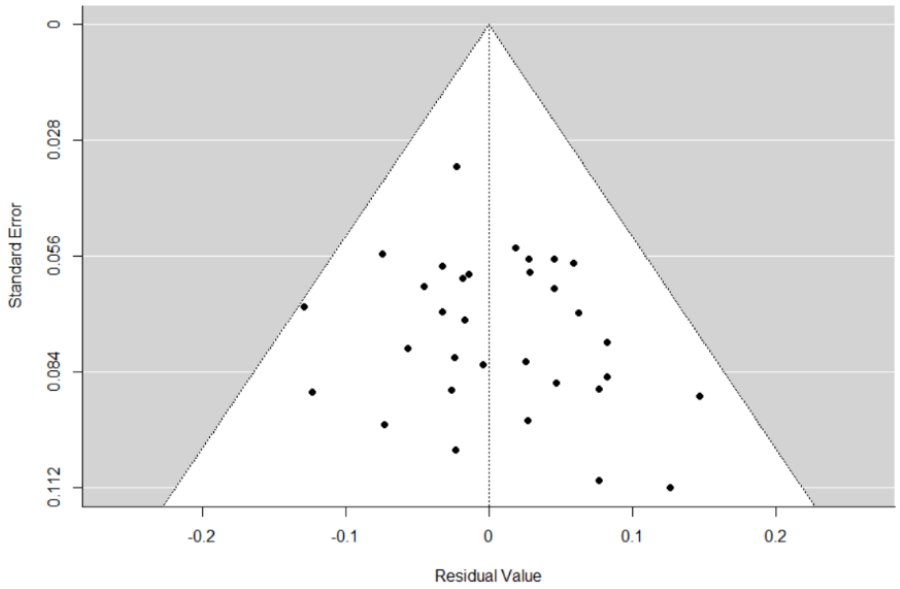
	Estimate	Standard error	t-value	p-value	95% CI Lower Bound	95% CI Upper Bound
Intercept	.23	.02	10.22	.51	0.19	0.28
See	.04	.03	1.51	.15	-0.01	0.10
Hear	.09	.03	3.38	<.01**	0.04	0.15
Hear judging	-.08	.03	-2.70	.01*	-0.14	-0.02
Silent	.03	.03	0.97	.34	-0.03	0.08

Note. *** indicates significance at the 0.1% level. ** indicates significance at the 1% level. * indicates significance at the 5% level.

[10] In addition to the three studies removed in Model 1 and 2, Broughton & Alexander FT1 was removed first then Kanthamani & Broughton Series 4, then Dalton Experiment 2 until no more studies were flagged as influential. These studies were removed to the same criteria for the previous models with standardized residuals exceeding ±2.

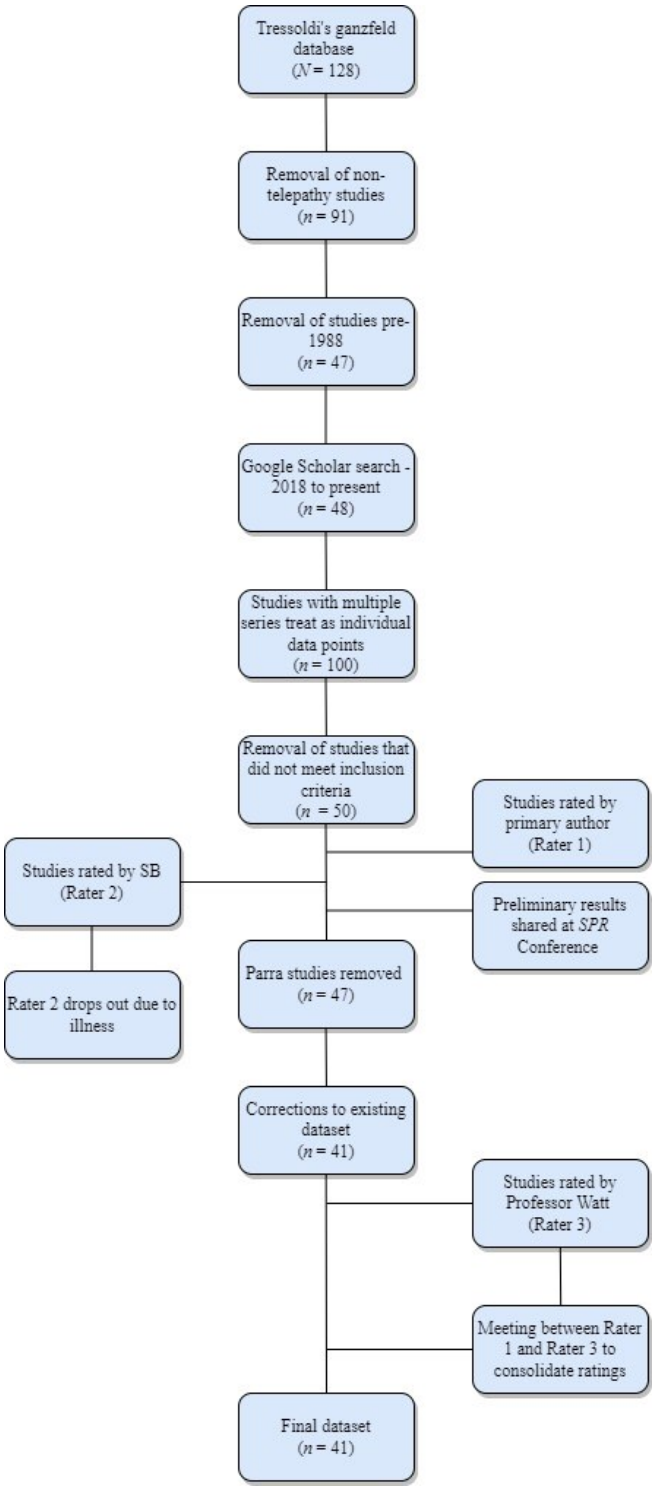
Supplemental Figure 4

Funnel plot for Model 1.2



Supplemental Figure 5

Flowchart of Study Selection



Supplemental Document 1

Rating Instructions

RATING INSTRUCTIONS

For each of the papers your task is to assess if they have certain characteristics present. If the characteristic is present then give it a 1, if not then a 0. Give a 1 if the characteristics are **explicitly** stated.

There are 5 characteristics to assess:

1. **Did the receiver see the sender's room?**
    a. *Some papers may say that both participants were shown the whole operation.*
2. **Did the sender hear the receiver produce their mentation (verbal report)?**
3. **Did the sender hear the receiver during the judging procedure?**
4. **Was the sender explicitly told to be silent?**
    a. *Some are told that any shouting/loud noises from sender's room would abort the session.*
5. **Did the experimenter review/allow additions to the mentation notes with the receiver, after the sending period?**
    a. *Some papers say they review the notes with the participant and allow for additions/changes.*

Use your judgement and common sense to assess if these characteristics were present, some will require re-reading and thinking but the main question is, "Is this clearly stated? Would I be able to run the exact same procedure given the detail in this paper?"

Important notes:

- Some shorter papers refer to other, already published papers and their procedures. Unless the author's state there were specific deviations from the previous design, you can give them the same ratings.
- Also, be careful for footnotes and procedural information outside of the 'Methods' section. It's worth skim reading all sections. Information about the study design may also be in the 'Participants'/'Procedure'/'Lab set-up'/'Design' parts of the paper (depending on how the paper is formatted).
- Some papers have multiple studies in them – you will be given a list of the studies of interest. However, this might require you to distinguish any differences in the procedures between the series, so it may take some deeper reading.

Supplemental Figure 6

Dataset

Study	ID	See	Hear	Judge	Silent	Review	HR	z-score	ES(h)	N participants	Trials
Alexander & Broughton (2001) - Telepathy sessions	125	1	0	0	1	1	36.00%	1.63	0.24	50	50
Broughton & Alexander (1997) - FT1[b]	101.1	1	1	0	1	1	24.00%	0	-0.023	50	50
Broughton & Alexander (1997) - FT2[a]	101.2	1	1	0	1	1	18.00%	-0.98	-0.171	50	50
Broughton & Alexander(1997) - EC	101.3	1	1	0	1	1	37.30%	1.81	0.267	51	51
Cardeña & Marcusson-Clavertz (2020)	143	1	0	0	0	1	22.86%	-0.1	-0.05	35	35
da Silva, Pilato & Hiraoka (2003) - Ganzfeld condition	113.1	1	1	1	0	1	33.33%	1.26	0.184	37	54
Dalton (1997) - Experiment 2[b]	133.2	1	1	0	1	0	53.00%	5.06	0.584	64	64
Dalton (1997) - Experiment 3	133.3	1	1	0	1	0	41.00%	2.75	0.343	64	64
Goulding, Westerlund, Parker & Wackermann (2004) - Receivers' judging[a]	127	1	1	1	1	1	14.00%	-1.88	-0.28	64	64
Honorton et al. (1990) - PRL Experienced 301	103.7	1	1	1	1	1	30.00%	0.65	0.112	25	50
Honorton et al. (1990) - PRL Experienced 302[a]	103.8	1	1	1	1	1	64.00%	4.26	0.807	25	25
Honorton et al. (1990) - PRL Novice 101	103.1	1	1	1	1	1	24.00%	0	-0.023	50	50
Honorton et al. (1990) - PRL Novice 102	103.2	1	1	1	1	1	36.00%	1.63	0.24	50	50
Honorton et al. (1990) - PRL Novice 103	103.3	1	1	1	1	1	30.00%	0.65	0.112	50	50
Kanthamani & Broughton (1994) - Series 1	105.1	0	1	0	0	1	19.40%	-0.52	-0.135	31	31
Kanthamani & Broughton (1994) - Series 3	105.3	0	0	0	0	1	20.00%	-0.55	-0.12	35	40
Kanthamani & Broughton (1994) - Series 4[b]	105.4	0	0	0	0	1	36.90%	2.08	0.259	65	65
Kanthamani & Broughton (1994) - Series 7	105.7	0	0	0	0	1	26.10%	0	0.025	27	46
Kanthamani & Broughton (1994) - Series 8	105.8	0	0	0	0	1	26.00%	0	0.023	16	50
Lau (2004) - Long-ganzfeld study 1	139.1	0	1	0	0	1	45.00%	1.8	0.423	20	20
Lau (2004) - Long-ganzfeld study 2	139.2	0	1	0	0	1	40.00%	1.29	0.322	20	20
Lau (2004) - Long-ganzfeld study 3	139.3	0	1	0	0	1	20.00%	-0.26	-0.12	20	20
Lau (2004) - Long-ganzfeld study 4	139.4	0	1	0	0	1	25.00%	0	0	20	20
Lau (2004) - Long-ganzfeld study 5	139.5	0	1	0	0	1	30.00%	0.26	0.112	20	20
Lau (2004) - Long-ganzfeld study 6	139.6	0	1	0	0	1	20.00%	-0.26	-0.12	20	20
Marcusson-Clavertz & Cardeña (2011)	134	1	0	0	0	0	27.00%	0	0.044	26	26
McDonough et al. (1994)	131	0	1	1	1	1	30.00%	0.26	0.112	20	20
Morris et al. (1993) - Cunningham	110.1	0	1	0	0	1	40.60%	1.84	0.334	32	32
Morris et al. (1995) - Honorton replication	107	1	1	1	1	1	34.00%	1.02	0.068	32	32
Morris, Summers & Yim (2003)	119	1	1	0	1	1	37.50%	1.643	0.271	40	40
Parker & Westerlund (1998) - Study 5	118.1	0	1	0	0	1	40.00%	1.27	0.261	30	30
Parker, Frederiksen & Johansson (1997) - Study 1	102.1	0	0	0	0	1	20.00%	-0.42	-0.12	30	30
Parker, Frederiksen & Johansson (1997) - Study 2	102.2	0	1	0	0	1	37.00%	1.27	0.261	30	30
Parker, Frederiksen & Johansson (1997) - Study 3	102.3	0	1	0	0	1	37.00%	1.27	0.261	30	30
Parker, Grams & Petterson (1998) - Study 4	108.4	0	1	0	0	1	47.00%	2.53	0.464	30	30
Roe, Holt & Simmonds (2003) - Human series	136	1	1	1	0	1	35.00%	1.28	0.219	40	40
Roe, McKenzie & Ali (2001)	135	0	0	0	0	1	20.83%	-0.24	-0.099	24	24
Roe, Sherwood & Holt (2004) - Sender condition	137.1	1	1	1	0	1	26.10%	0.12	0.025	23	23
Simmonds-Moore & Holt (2007) - Ganzfeld condition	141.1	1	1	1	0	1	23.10%	0	-0.044	26	26
Smith & Savva (2008)	114	1	1	0	0	1	34.20%	2.16	0.202	114	114
Wright & Parker (2003) - One session	129.1	1	1	0	0	0	44.73%	2.62	0.418	10	38
Totals (factor, participants, trials)		**22**	**32**	**12**	**15**	**37**	**31.85%[c]**	**0.91[c]**	**0.139[c]**	**1496**	**1624**

Note. [a] indicates the three studies that were removed from both Models 1, 1.1 and 2 due to influence [b] indicates the three studies removed from Model 1.2 due to influence [c] is the column mean.

Appendix B

Chapter 4 Supplemental Materials

This appendix contains information about each examined dataset in the umbrella review and other key information, including database search information, evidence for the Which coding and extraction of statistics for Supplemental Table 1.

Supplemental Table 1

Summary statistics of each meta-analysis in umbrella review, sorted by year of publication

Meta-analysis	N studies	N session	Hits	Hit rate (%) [CI]	*z*-score	*z*-score CI	ES value[a]	ES CI
Honorton (1985)	28	835	302	36.17	1.25	.76 - ?	6.60 (Stouffer's Z)	
Hyman (1985)	**36**			**34.00**	**1.04**		**5.98**	
Honorton (1990)	11	355	122	34.40 [30-39]	3.89		0.2	
Bem & Honorton (1994)	**10**	**329**	**106**	**32.00 [30-35]**	**2.89**		**0.59**	**.53-.64**
Honorton (1995)	73	4155	1330*	32.20	5.74		0.16	.06-.26
Milton (1997)	**42**	**1572***	**459***	**29.20**			**2.49 (Stouffer's Z)**	
Honorton (1998)		221						
Milton & Wiseman (1999)	**30**	**1198**	**327***	**27.30**	**0.13**		**0.013**	
Palmer & Broughton (2000)	10	463	151*	32.60	1.11		0.165	
Bem et al. (2001)	**10**	**463**	**170***	**36.70**	**1.26**		**0.17**	
Storm & Ertel (2001)	11	405	128*	31.60*	1.04		0.222	
Schlitz (2003)		**2904***	**929**	**32.00**	**8.75**			

Radin (2006)	88	3145	1008	32.00			0.16	.14-.18
Utts et al. (2010)	**56**	**2124**	**709**	**33.40**	**8.92**			
Storm et al. (2010b)	30	1520*	490*	32.24*	1.16		0.152 6.34 (Stouffer's *Z)*	
Williams (2011)	**59**	**2832**	**878**	**31.00**	**7.37**			
Rouder et al. (2013)	19*	1058*	321*	30.34*				
Storm et al. (2013)	**29**	**1520***	**490***	**32.24**	**1.01**		**0.14**	
Storm & Tressoldi (2020)	9	471	146	31.00	0.854	0.15-1.56	0.119	.003-.235
Pooley et al. (2023)	**41**	**1624**	**520**	**32.00**	**0.91**		**5.81 (Stouffer's *Z*)**	
Tressoldi & Storm (2023)	113	4841	1520	31.40			.099	.05-.14

Note: * indicates statistics are not reported in the article but deduced and calculated from the article information. Blank cells indicate missing information that was either not presented in the original article nor could it be deduced or calculated. [a] We report ES as identified in Table 2 but for studies which used multiple ES, we chose a common one to allow for comparison.

1. Honorton (1985)

We use the (only) database, as defined, in this article. Honorton produced this response paper to Hyman (1985). Analysis of subset (28/42) studies assessed by Hyman, that reported direct hits. Table 4 hits and hit rate calculated from Table A1 of meta-analysis. DARE2 and DARE3 are no, given Honorton stated he wanted to re-analyse the database he provided Hyman.

Database search strategy

Year range search	Electronic database search	Manual database search
1974-1981	N/A	28/42 papers considered by Hyman (1985): Journals (unidentified); Parapsychological Association convention proceedings in Research in Parapsychology and monographs.

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Other combination	Manual ganzfeld only	Both	Any	Yes
Clairvoyance and telepathy p.77-78	Pre-autoganzfeld due to era	Mostly unselected but Palmer et al. (1979) used meditators	p.69-70	Unpublished honours thesis p.83

2. **Hyman (1985)**

We use the (only) database, as defined, in this article. Hyman was tasked to assess parapsychology, given a database of 42 studies compiled by Honorton. 36/42 studies as defined by Hyman, Step 4 p.8. Used 36 studies which provided data to calculate common effect size and significance test. Search strategy general but objective was to analyse all known studies within a time range, also contacted an expert (DARE2). Include everything as inclusion strategy (DARE3).

Database search strategy

Year range search	Electronic database search	Manual database search

1974-1981	N/A	Journal of the American Society for Psychical Research; European Journal of Parapsychology; Journal of the Society for Psychical Research; published monographs; abstracts/papers at Parapsychological Association Annual Conferences

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All ESP	Manual ganzfeld	Both	Any	Yes
p.10 defines telepathy studies included and clairvoyance and precognition studies p.47-49	Pre-autoganzfeld due to era	Suspect mostly unselected but transcendental meditators in reference list	p.27	p.38 and reference list

3. **Honorton et al. (1990)**

We use the (only) database, as defined, in this article. This is a meta-analysis which only includes the 11 studies conducted at the PRL labs and no literature search was conducted, hence DARE2 and DARE3 are not met.

Database search strategy

Year range search	Electronic database search	Manual database search
All PRL autoganzfeld studies, 1983-1989	N/A	All PRL autoganzfeld studies

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Other combination	Auto-ganzfeld only	Both	Automated	Yes
Telepathy and clairvoyance series	Developing the new autoganzfeld method	Novice series (unselected) and other with low beliefs, a mixture but no formal screening.	Quasi-random given human involvement but computer software had random number generator	First published report of all 11 PRL studies (thus using studies which have not been previously published)

4. **Bem and Honorton (1994)**

We use the only database defined in this article, which is the same database as Honorton et al. (1990). But this one is 10 studies rather than 11 as study 302 was analysed separately by Bem and Honorton. Again, the authors did not conduct a literature search for the studies, Hence DARE2 and 3 are not met. Honorton et al. (1990) has 'Combination' for the Study Design factor, but Bem and Honorton does not as nowhere is it specified in the B\&H analysis of any clairvoyance sessions, although series 103 in Honorton et al. (1990) is stated to have a clairvoyance option.

Database search strategy

Year range search	Electronic database search	Manual database search
All PRL autoganzfeld studies, 1983-1989	N/A	All PRL autoganzfeld studies

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Telepathy	Auto-ganzfeld only	Both	Automated	No
Receiver and sender details in methods section p.10	Developing the new autoganzfeld method, p.9	p.10 in *Experimental Studies* section defines novices and experienced, plus Juilliard sample	Noise-based random number generator, p.10	Footnote 4 defines this dataset previously published by Honorton et al. (1990)

5. **Honorton (1995)**

We use the only database, as defined, in this article. This database evaluates the role the presence of a sender (or none) has on session outcome.

Database search strategy

Year range search	Electronic database search	Manual database search
1974-1981	N/A	All English language published in parapsychology literature (unspecified)

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Other combination	Mixed	Both	Any	Yes
No study compared sender/no sender within studies p.135	Date range of studies searched and reference list has autoganzfeld and manual series studies.	Not explicitly stated; studies in reference list have selected and unselected participants	Not explicitly stated; studies in reference list mix of manual and autoganzfeld studies	Defined in method, p.133

6. **Milton (1997)**

We assessed the 42 studies which used the ganzfeld method. We report the rounded-up direct hits data for ganzfeld studies in Table 1 of the article, as reported in Table 4 (above). At this time, direct hits were the most commonly used and published. Milton's Appendix states 44 ganzfeld studies but she only had the full data for 42 ganzfeld sessions to calculate statistics. The *N* sessions and hit rate in Table 4 were calculated by hand using the information provided in Milton's Appendix.

Database search strategy

Year range search	Electronic database search	Manual database search
1992-1996(?)	N/A	European Journal of Parapsychology; Journal of the American Society for Psychical Research; Journal of Parapsychology; Journal of the Society for Psychical Research; Proceedings of Parapsychological Association Conference; B&H (1994)

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	No
Reference list of studies included in Appendix	Mainly autoganzfeld but Kanthamani manual series included (Appendix)	Reference list of studies included in Appendix	Not specified but given studies included likely a mix	Retrieved published studies only p.230, but some unpublished data (sum of ranks calculations)

7. **Honorton et al. (1998)**

We coded the 'A New Confirmation' database from page 268, which is a subset of the PRL participants (Honorton et al., 1990) whom had extraversion data available (221/241). The earlier database in the Honorton et al. (1998) paper was not coded as there is no mention of the ganzfeld paradigm in the free-response studies. N studies unknown even though 11 studies in PRL database but not specified which studies the 221 out of the 241 are taken from.

Database search strategy

Year range search	Electronic database search	Manual database search
N/A but only includes PRL dataset which was conducted 1983-1989	N/A	PRL studies only with extraversion data

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Combination	Autoganzfeld only	Both	Automated	No
Not explicitly stated; PRL were clairvoyance and telepathy studies	Not explicitly stated but PRL was first to use autoganzfeld	PRL had a mixture of novices and selected characteristics	Not explicit but PRL was automatic target randomisation	Analysis is using a subset of already published PRL studies

8. **Milton and Wiseman (1999)**

We coded the (only) database in this article. The asterisk in Table 4 is for the number of hits which are extracted from Storm et al. (2010b).

Database search strategy

Year range search	Electronic database search	Manual database search
1987-1997	N/A	Main parapsychology journals (unspecified) and Proceedings of the Annual Parapsychological Association Conference

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	No
No specific statement but reference list has all three types	Manual (Kanthamani series) and autoganzfeld studies included	Not specified but Willin studies had no selection criteria whilst Honorton (1997) is four-factor model	Not specified but most are autoganzfeld studies with automated randomisation but Stanford and Frank use random number tables	"No attempt was made to find unpublished studies" p.388

9. **Palmer and Broughton (2000)**

We coded the 10 new studies collated by the authors in their conference proceedings paper (the final, published version of the meta-analysis is Bem et al., 2001). Palmer and Broughton added 10 new studies to the 30 studies from the Milton and Wiseman (1999) database – we are interested in the new database. This meta-analysis was a response to the non-significant findings of the Milton and Wiseman analysis. We hand calculated the number of sessions from Table 1 in the article.

Database search strategy

Year range search	Electronic database search	Manual database search
1996-1999	N/A	European Journal of Parapsychology; Journal of the American Society for Psychical Research; Journal of Parapsychology; Journal of Scientific Exploration; Journal of the Society for Psychical Research; Proceedings of the Parapsychological Association

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Autoganzfeld	Both	Automated	No
"All studies described as using some sort of ganzfeld procedure were included" p. 226	Not explicitly stated but year range search is autoganzfeld era	Standardness ratings give more weight for selected, suggesting also accept unselected	Not explicitly stated but autoganzfeld era	Not specified in methods but also no unpublished studies in reference list

10. **Bem et al. (2001)**

We coded the 10 new studies collated by the authors. This meta-analysis was in response to the non-significant results reported by Milton and Wiseman (1999). This is the same database as their conference proceeding (Palmer & Broughton, 2000) but we use both versions of this database to see if there are any differences. Again, the number of sessions calculated from totalling values in Table 1 of the article. Note the difference in hits between the two versions of this database due different studies being asterisked (10 new studies) between the Palmer and Broughton (2000) and Bem et al. (2001) articles, namely the Parker et al. (1997) studies in the Palmer (2000) not asterisked in the Bem et al. (2001) paper, but three Parker and Westerlund (1998) series being asterisked instead.

Database search strategy

Year range search	Electronic database search	Manual database search

Not specified	N/A	6 major publications for parapsychological research (unspecified)

Which factor rationale

Study design	**Study mechanism**	**Participant type**	**Randomisation method**	**Unpublished studies**
Combination	Autoganzfeld	Both	Automated	No
Not specified but the 10 new studies are telepathy and clairvoyance only	Not specified but given the reference list of the 10 new studies are autoganzfeld era	Not specified but quality ratings give more points for selected versus unselected suggesting accept both	Not specified but new 10 studies are autoganzfeld era	Method section states "six major publications" suggesting no search for unpublished studies

11. **Storm and Ertel (2001)**

We coded the Storm and Ertel database of 11 'previously overlooked studies' which were conducted between 1982-1986, pre-*Joint Communiqué*. Little detail is given about the databases they searched. This article was a response to the non-significant findings reported by Milton and Wiseman (1999). For the statistics, we report the non-quality weighted results due to the different quality ratings used across the different meta-analyses.

Database search strategy

Year range search	Electronic database search	Manual database search
1982-1986	Not specified	Not specified

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Manual only	Both	Any	No
Not explicitly stated but reference list has all three designs and aim of meta-analysis to assess "unified domain of ganzfeld data" p.426 suggesting all design include	States that they are looking at manual, pre-autoganzfeld studies only p.427	Quality criteria gives more weight to selected participants versus unselected p.427	Not stated but quality assessment says "randomized targets were used" p.427 suggesting any type included	Not explicitly stated but no unpublished studies in reference list

12. **Schlitz and Radin (2003)**

Non-sensory access to information: the ganzfeld studies chapter in Healing, Intention and Energy Medicine. We coded the meta-analysis reported in Chapter 7 of this book, chapter was compiled in late 2001. Given the format (popular science book) there is very little detail about how the meta-analysis was constructed but this reference is cited within the parapsychology literature, hence its inclusion. The authors conduct a new database by combining datasets from meta-analyses published in 1985, 1994, 1999 and 2001. Which factors heavily deduced for this meta-analysis given the lack of information.

Database search strategy

Year range search	Electronic database search	Manual database search
"All known ganzfeld experiments" up until late 2001	Not specified	Not specified

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	Not specified
Not specified but given “all known ganzfeld experiments” for their search	Heavily deduced from the year range searched	Heavily deduced from year range search as spans manual and autoganzfeld time periods	Not specified but heavily deduced given time range searched	Harder to deduce as no detail provided on sources

13. **Radin (2006)**

Conscious Psi chapter in Entangled Minds. We coded the meta-analysis reported in pages 120-121 and chapter endnotes. Again, as it is a book chapter, little information provided about how the database was compiled and Which factors are heavily deduced.

Database search strategy

Year range search	Electronic database search	Manual database search
1974-2004	Google search on November 27, 2004	Not specified

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	Not specified
Deduced from footnotes and main text given date range and only early studies being excluded due to hit vs miss type analysis	Not explicitly stated but year range includes manual and autoganzfeld era	Heavily deduced from date range	Heavily deduced from date range	Can't deduce from information provided

14. **Utts et al. (2010)**

This is a conference workshop tool to demonstrate Bayesian statistics using ganzfeld data. This meta-analysis is an ad-hoc combination of studies to create a dataset but states the included studies are those which "met criteria for methodological rigor and adherence to standard ganzfeld procedures". Utts et al. (2010) state they include 16 studies from Table 2 in Dawson (1991) -- these are studies from Honorton's (1985) meta-analysis, excluding 8 from Honorton's full database which had unresolved allegations of methodological flaws. The Utts database is included in the systematic review, even though it is a workshop tool, as it is cited in the within the parapsychology literature. Given the database is handpicked, we extracted most information for the Which factors from the primary studies included. We extracted the frequentist statistics for Table 4 to allow comparison with the majority of our database.

Database search strategy

Year range search	Electronic database search	Manual database search

Not specified	Not specified	Handpicked: 16 from Dawson (1991) Table 2, all 11 in Table 1 of Bem and Honorton (1994), 29 studies with "standardness score" more than 4 in Table 1 of Bem et al. (2001)

Which factor rationale

Study design	**Study mechanism**	**Participant type**	**Randomisation method**	**Unpublished studies**
Combination	Mixed	Both	Any	Not specified
Handpicked databases only include telepathy and clairvoyance studies	Date range of the studies handpicked include manual and autoganzfeld studies	Date range of studies handpicked include selected and unselected series	Handpicked studies from previous meta-analyses include manual and automated randomisation methods	Handpicked studies are all from published meta-analyses but the individual studies from Table 2 in Dawson (1991) are not cited so cannot confirm if unpublished studies included

15. **Storm et al. (2010b)**

We coded Group A (ganzfeld) set of studies and coded the Which factors from the methods section on p.474-475. The non-homogeneous dataset ($N = 30$) was extracted and is reported in Table 4, with the number of hits, session and hit rate calculated from the Appendix of the Storm paper. Storm et al. (2010b) compiled 30 studies but later removed one study (Dalton, 1997) to create their homogeneous dataset.

Database search strategy

Year range search	Electronic database search	Manual database search
1997-2008	PsycINFO; PsycARTICLES; CINAHL; Medline; Web of Science; Lexscien; InformIT	Journal of Parapsychology; European Journal of Parapsychology; Journal of the Society for Psychical Research; Journal of the American Society for Psychical Research; Proceedings of the Annual Convention of the Parapsychological Association; Journal of Scientific Exploration; Australian Journal of Parapsychology; Journal of Cognitive Neuroscience; Dreaming

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	Yes
Stated in methods p.474	Time period of studies, also include Kanthamani manual series	Directly compare selected and unselected participants	As stated in methods section p.474, random number generator in a computer or table of random numbers (manual)	Include Lau (2004) unpublished Master's thesis (reference list)

16. Williams (2011)

As stated, this article is often cited in the parapsychology ganzfeld literature but the author, Williams, explicitly states it is not a meta-analysis and should not be treated as such, p. 648. Given it is frequently cited in the ganzfeld literature as further support of replicable evidence, we include this article in our review and coded like any other. Like the Utts et al. (2010) workshop tool, the database created by Williams is hand-picked, rather than having a formal literature search procedure. We code the only database in this article.

Database search strategy

Year range search	Electronic database search	Manual database search
1987-2008 Aggregated post-PRL studies from Bem et al., 2001; Milton and Wiseman, 1999 and Storm et al., 2010b	N/A	3 previous meta-analyses combined

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	Yes
Not explicit but Appendix 1 and aggregating other databases suggesting no exclusion of certain designs	Not explicit but Appendix 1 includes Kanthamani manual series and numerous autoganzfeld (post-PRL) studies	Not explicit but Appendix 1 includes primary studies involving selected and unselected participants	Not explicit but no specific exclusion of certain randomisation methods	Lau (2004) unpublished master's thesis in reference list, from Storm et al., 2010b

17. Rouder et al. (2013)

This a response article to Storm et al. (2010b). Rouder and colleagues create three datasets, the first is the same as Storm et al. (2010b) minus May (2007), Revised Set 1 is automated studies only and Revised Set 2 is Revised Set 1 plus the omitted conditions from two studies from Storm et al. We focus on Revised Set 1 as it is the new database constructed by the authors and more information is provided in their appendix. The statistics in Table 4 are manually extracted as Rouder et al. aggregated all three categories from Storm et al.'s (2010b) analysis and thus, should be taken with caution.

Database search strategy

Year range search	Electronic database search	Manual database search
1997-2008	N/A No search conducted, adjusted Storm et al. (2010b) database	N/A No search conducted, adjusted Storm et al. (2010b) database

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Autoganzfeld	Both	Automated	Yes
Same database as Storm et al. (2010b), no comment on excluding certain designs	Revised Set 1 excludes all manual studies p.243	Same database as Storm et al. (2010b), no comment on excluding participant types	Not explicit but since only analysing autoganzfeld studies, extrapolate that automated randomisation only	Lau (2004) included; footnote 2 p. 242

18. Storm et al. (2013)

This is a response article to Rouder et al. (2013) using their original database from 2010b, with some minor updates but no new database search was conducted. The authors once again split their analysis into three categories, as per their 2010b analysis. We coded the homogenous ganzfeld data (Category 1) as they did not provide statistics for the nonhomogeneous data. However, for this analysis, studies 7 and 11 were corrected as described p. 251. Statistics in Table 4 are extracted from Table 2 and footnotes of Storm et al. (2013) and hand calculated from the Storm 2010b dataset. The footnotes in the 2013 article provide corrected hits and sessions, hence the different hits and session numbers between Storm et al. (2010b) and Storm et al. (2013) in Table 4.

Database search strategy

Year range search	Electronic database search	Manual database search
1997-2008 (same as Storm et al., 2010)	N/A Response article using same Storm et al. (2010b) database	N/A Response article using same Storm et al. (2010b) database

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	Yes
Using Storm et al. (2010b) database which included all	Using Storm et al. (2010b) database which included both	Using Storm et al. (2010b) database which accepted both	Directly compare manual and automated randomisation p.250	Not explicit but if same Storm et al. (2010b) dataset then includes unpublished studies No comment about excluding unpublished

19. Storm and Tressoldi (2020)

We coded the updated database (not the amalgamated). This meta-analysis provides an update to their 2010b analysis to assess if their results are replicated.

Database search strategy

Year range search	Electronic database search	Manual database search
2009-2018	PsycINFO; PsycARTICLES; Medline; Web of Science; Lexscien; Informit	Australian Journal of Parapsychology; Consciousness and Cognition; Dreaming; European Journal of Parapsychology; Europe's Journal of Psychology; Journal of Cognitive Neuroscience; Journal of Parapsychology; Journal of Scientific Exploration; Journal of the Society for Psychical Research; Proceedings of the Annual Convention of the Parapsychological Association

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies

All	Autoganzfeld	Both	Automated only	No
p. 196 “the three major psi modalities”	Not explicit but given the date range searched	p. 196 and directly comparing selected and unselected	p. 199	Not explicitly stated but no unpublished studies in reference list

20. Pooley et al. (2023)

We code the only database in this meta-analysis. The extracted statistics in Table 4 are for all 41 studies in the descriptive statistics section p. 54 of the artle, due to two different statistical models being used later in the analysis. This article conducts a meta-regression, which focused on five characteristics of telepathy studies.

Database search strategy

Year range search	Electronic database search	Manual database search
1988-2021	Google Scholar	Reference lists of included studies

Which factor rationale

Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
Telepathy	Mixed	Both	Any	Yes
Focus of the meta-analysis	Time span includes pre-autoganzfeld	Not explicitly stated but studies not	Not explicitly stated but studies not	Lau (2004) and Dalton (1997) unpublished theses included

	included/excluded based on participant type	included/excluded due to randomisation methods

21. Tressoldi and Storm (2023) – Stage 2

We code the only database in this meta-analysis. As this is a public peer-review, we coded the most recent one at the time of the database searching.

Database search

Year range search	Electronic database search	Manual database search
1974-2020	Google Scholar; PubMed; Scopus	Database aggregated by Storm and Tressoldi (2020); specialised journals (unspecified)

Which factor rationale

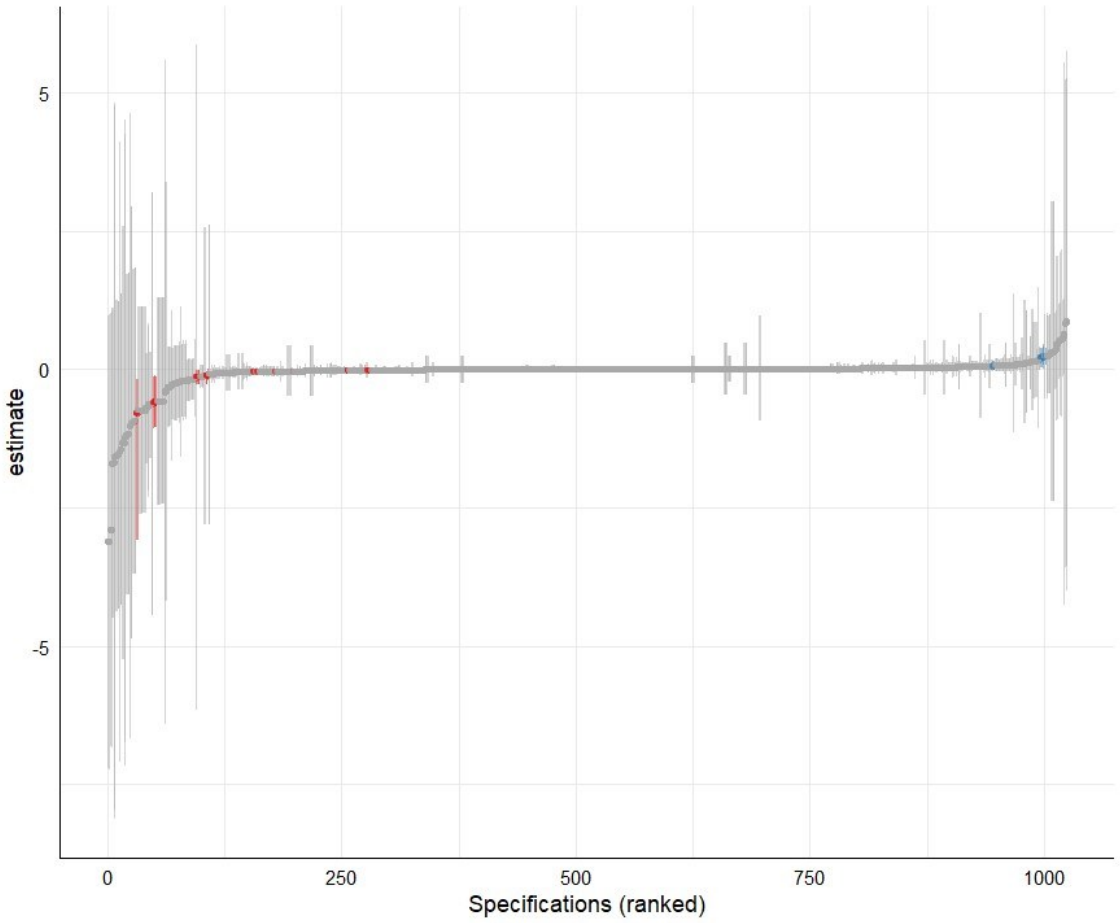
Study design	Study mechanism	Participant type	Randomisation method	Unpublished studies
All	Mixed	Both	Any	No
Coded based on design type	Not explicitly stated but the date range is from early ganzfeld to current	Variables coding section	Studies inclusion criteria	Retrieval method implies published only

Appendix C

Chapter 6 Supplemental Materials

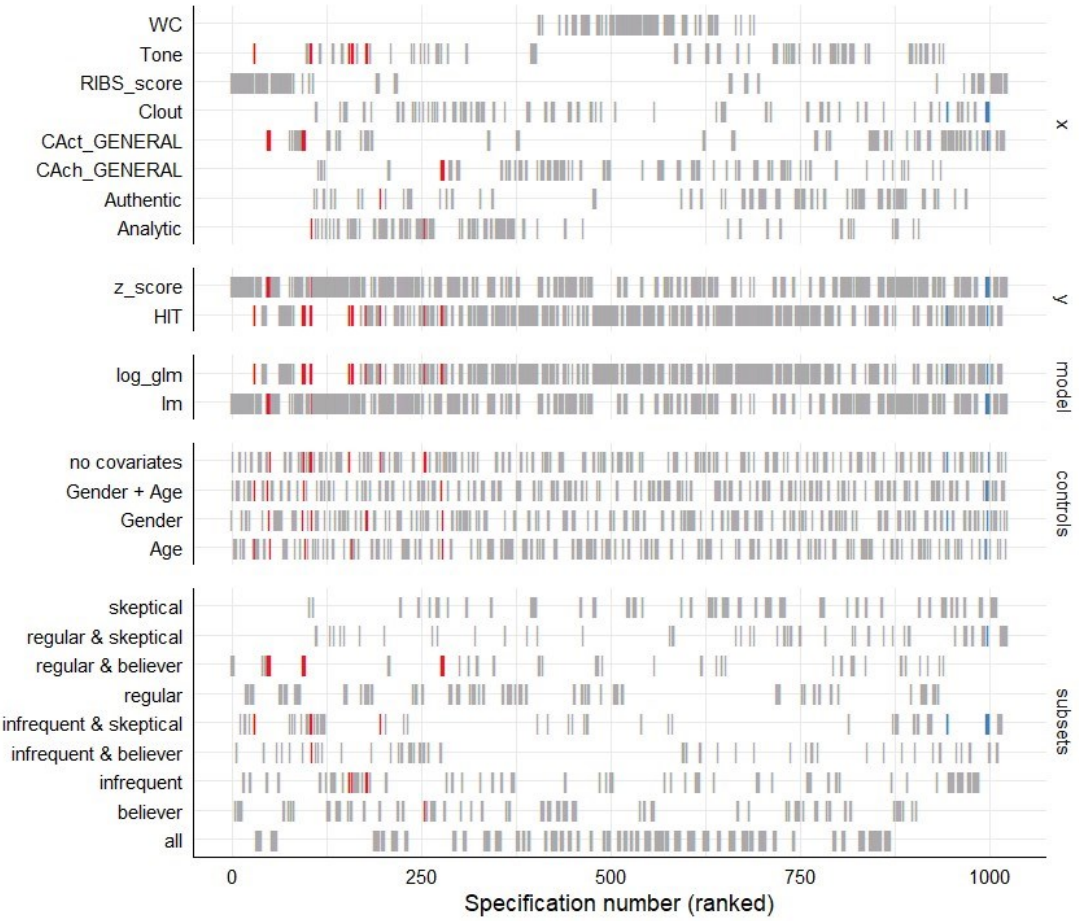
Supplemental Figure 1

Panel A of Figure 6.2



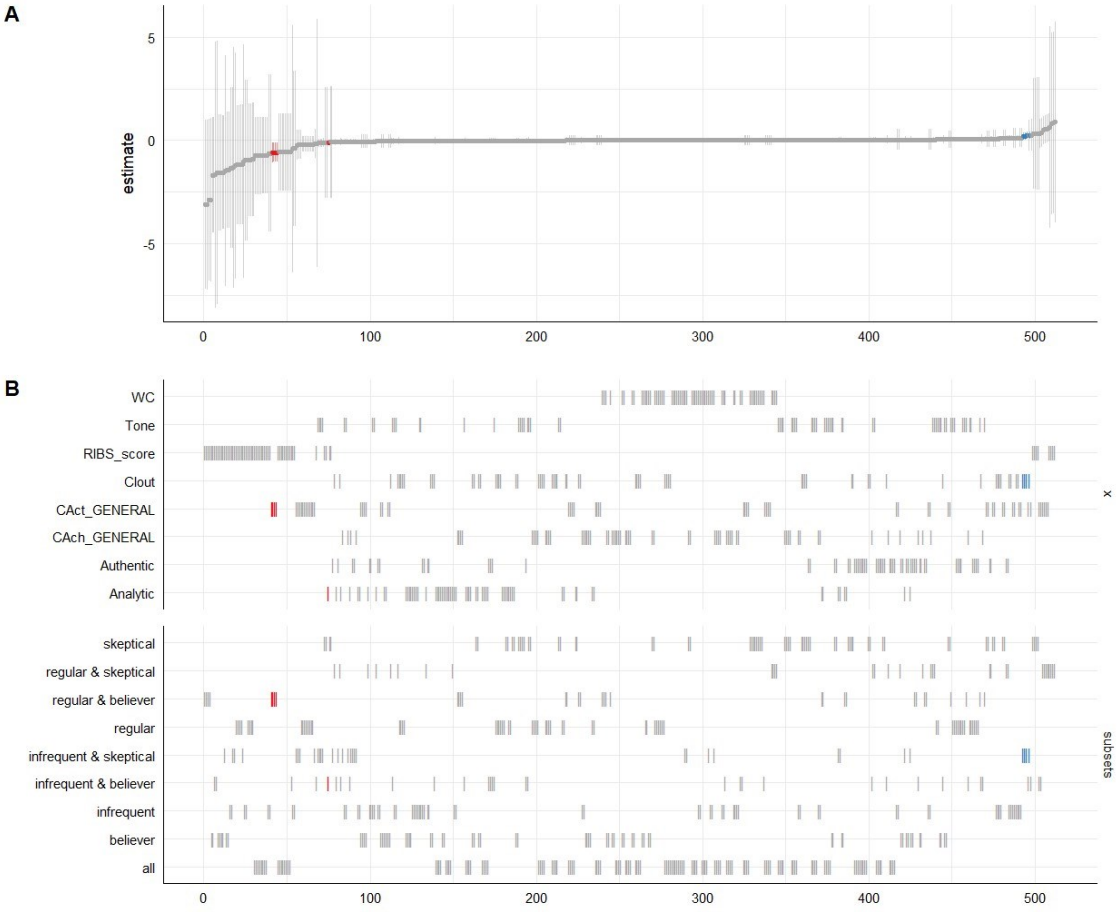
Supplemental Figure 2

Panel B of Figure 6.2



Supplemental Figure 3

Results of Second Specification Curve Analysis, presenting the Independent Variables and Subsets only



Appendix D

Chapter 7 Supplemental Materials

Supplemental Table 1

Descriptive Statistics of LIWC2015 Analysis of KPU Ganzfeld Mentations

				95% CI Mean				
	Valid	**Missing**	***M***	***UL***	***LL***	***SD***	***Min.***	***Max.***
Word Count	251	49	505.120	567.913	442.737	500.048	2.000	2942.000
Analytic	251	49	68.663	72.046	65.228	27.779	4.310	99.000
Clout	251	49	32.306	34.826	30.104	19.793	1.000	92.330
Authentic	251	49	69.615	72.913	66.108	28.892	1.000	99.000
Tone	251	49	41.752	44.640	39.081	23.261	1.000	99.000
WPS	251	49	8.213	8.749	7.642	4.542	1.000	29.600
Sixltr	251	49	14.786	15.589	14.036	6.468	0.000	50.000
Dic	251	49	84.625	86.035	83.144	11.787	0.000	100.000
function	251	49	52.503	54.019	50.792	13.314	0.000	69.840
pronoun	251	49	12.986	13.703	12.221	6.100	0.000	27.470
ppron	251	49	7.568	8.192	7.023	4.791	0.000	25.000
i	251	49	6.135	6.687	5.586	4.528	0.000	25.000
we	251	49	0.203	0.260	0.146	0.468	0.000	4.320
you	251	49	0.414	0.511	0.333	0.736	0.000	6.520
shehe	251	49	0.406	0.516	0.300	0.884	0.000	7.490
they	251	49	0.410	0.484	0.339	0.576	0.000	3.230
ipron	251	49	5.379	5.742	4.972	3.169	0.000	16.220
article	251	49	10.581	11.214	9.944	5.240	0.000	50.000
prep	251	49	13.202	13.766	12.622	4.525	0.000	26.030
auxverb	251	49	7.747	8.226	7.213	4.040	0.000	15.530
adverb	251	49	5.912	6.329	5.489	3.400	0.000	16.670
conj	251	49	6.165	6.560	5.789	3.251	0.000	14.810
negate	251	49	1.232	1.410	1.073	1.391	0.000	14.290
verb	251	49	16.344	17.033	15.572	5.769	0.000	34.620
adj	251	49	4.493	4.859	4.165	2.813	0.000	29.410
compare	251	49	4.095	4.500	3.679	3.367	0.000	18.550
interrog	251	49	0.737	0.832	0.642	0.746	0.000	3.880
number	251	49	1.102	1.942	0.586	6.397	0.000	100.000
quant	251	49	1.820	1.997	1.664	1.355	0.000	8.330
affect	251	49	2.432	2.681	2.206	2.010	0.000	17.650

	Valid	Missing	*M*	95% CI Mean *UL*	95% CI Mean *LL*	*SD*	*Min.*	*Max.*
posemo	251	49	1.625	1.805	1.443	1.491	0.000	11.760
negemo	251	49	0.795	0.917	0.674	1.012	0.000	6.580
anx	251	49	0.210	0.265	0.157	0.443	0.000	2.780
anger	251	49	0.148	0.194	0.111	0.321	0.000	2.630
sad	251	49	0.207	0.288	0.145	0.567	0.000	5.880
social	251	49	4.097	4.439	3.781	2.751	0.000	15.540
family	251	49	0.191	0.243	0.144	0.403	0.000	2.610
friend	251	49	0.157	0.208	0.115	0.382	0.000	3.450
female	251	49	0.396	0.530	0.297	0.930	0.000	10.000
male	251	49	0.477	0.581	0.371	0.887	0.000	8.440
cogproc	251	49	10.365	11.018	9.728	5.395	0.000	25.000
insight	251	49	2.608	2.859	2.364	2.007	0.000	11.110
cause	251	49	0.703	0.798	0.614	0.730	0.000	4.740
discrep	251	49	0.593	0.700	0.491	0.827	0.000	6.250
tentat	251	49	3.966	4.288	3.614	2.739	0.000	16.020
certain	251	49	0.688	0.785	0.606	0.755	0.000	6.250
differ	251	49	2.927	3.178	2.670	2.054	0.000	9.380
percept	251	49	8.711	9.330	8.144	4.992	0.000	41.180
see	251	49	5.414	5.971	4.936	4.154	0.000	34.550
hear	251	49	0.921	1.165	0.734	1.715	0.000	22.220
feel	251	49	2.111	2.459	1.821	2.557	0.000	23.530
bio	251	49	3.294	3.753	2.908	3.428	0.000	28.570
body	251	49	1.729	1.998	1.486	2.195	0.000	17.240
health	251	49	0.414	0.583	0.280	1.253	0.000	14.290
sexual	251	49	0.016	0.026	0.008	0.073	0.000	0.730
ingest	251	49	1.170	1.391	0.959	1.835	0.000	14.290
drives	251	49	3.233	3.479	2.987	1.993	0.000	10.110
affiliation	251	49	0.668	0.784	0.553	0.922	0.000	5.760
achieve	251	49	0.425	0.494	0.356	0.527	0.000	2.610
power	251	49	1.597	1.773	1.443	1.304	0.000	6.820
reward	251	49	0.555	0.640	0.476	0.682	0.000	4.100
risk	251	49	0.158	0.204	0.119	0.354	0.000	2.780
focuspast	251	49	1.691	1.897	1.503	1.533	0.000	8.330
focuspresent	251	49	12.099	12.780	11.399	5.640	0.000	25.000
focusfuture	251	49	0.913	1.028	0.807	0.887	0.000	6.150
relativ	251	49	15.036	15.666	14.466	4.826	0.000	29.790
motion	251	49	2.845	3.093	2.610	1.926	0.000	15.170
space	251	49	9.056	9.543	8.605	3.766	0.000	23.530
time	251	49	3.483	3.776	3.217	2.163	0.000	19.050
work	251	49	0.544	0.651	0.445	0.827	0.000	5.660
leisure	251	49	1.923	2.183	1.670	2.117	0.000	14.290
home	251	49	0.843	1.044	0.679	1.424	0.000	14.290

	Valid	Missing	*M*	95% CI Mean *UL*	95% CI Mean *LL*	*SD*	*Min.*	*Max.*
money	251	49	0.111	0.147	0.077	0.286	0.000	2.200
relig	251	49	0.173	0.226	0.124	0.413	0.000	3.450
death	251	49	0.110	0.146	0.078	0.294	0.000	2.520
informal	251	49	0.692	0.892	0.533	1.406	0.000	11.540
swear	251	49	0.036	0.064	0.014	0.211	0.000	2.630
netspeak	251	49	0.051	0.078	0.029	0.198	0.000	1.750
assent	251	49	0.192	0.254	0.139	0.467	0.000	3.340
nonflu	251	49	0.410	0.586	0.267	1.290	0.000	11.540
filler	251	49	0.010	0.018	0.004	0.058	0.000	0.790

Note. 95% *CIs* for mean estimated with 1,000 bootstraps.

Supplemental Table 2

Descriptive Statistics of LIWC2015 Analysis of KPU Ganzfeld Mentations by Session Outcome

	Hit	Valid	Missing	*M*	95% CI Mean *UL*	95% CI Mean *LL*	*SD*	*Min*	*Max*
Word Count	0	170	36	496.494	577.556	421.735	523.631	2.000	2942.000
Word Count	1	81	13	523.222	619.503	437.289	449.093	7.000	1763.000
Analytic	0	170	36	70.178	74.452	66.011	27.718	4.540	99.000
Analytic	1	81	13	65.483	71.351	59.488	27.808	4.310	99.000
Clout	0	170	36	33.268	36.052	30.274	20.341	1.000	82.830
Clout	1	81	13	30.288	34.437	26.308	18.550	2.860	92.330
Authentic	0	170	36	67.831	72.088	63.640	29.506	1.000	99.000
Authentic	1	81	13	73.359	79.296	67.358	27.358	1.310	99.000
Tone	0	170	36	41.599	44.798	38.242	23.525	1.000	99.000
Tone	1	81	13	42.071	46.895	37.056	22.836	1.000	98.900
WPS	0	170	36	8.165	8.843	7.547	4.490	1.190	29.600
WPS	1	81	13	8.315	9.409	7.386	4.675	1.000	22.320
Sixltr	0	170	36	14.823	15.765	13.863	6.480	0.000	50.000
Sixltr	1	81	13	14.708	16.123	13.335	6.482	0.000	42.860
Dic	0	170	36	84.230	85.931	82.370	12.058	0.000	100.000
Dic	1	81	13	85.452	87.749	82.741	11.223	30.190	96.940
function	0	170	36	52.085	53.939	50.056	13.579	0.000	69.840
function	1	81	13	53.381	56.024	50.590	12.777	0.000	67.870
pronoun	0	170	36	12.746	13.609	11.813	6.112	0.000	27.470
pronoun	1	81	13	13.488	14.732	12.122	6.082	0.000	23.810
ppron	0	170	36	7.441	8.265	6.768	4.853	0.000	25.000

					95% CI Mean				
	Hit	**Valid**	**Missing**	***M***	***UL***	***LL***	***SD***	***Min***	***Max***
ppron	1	81	13	7.835	8.851	6.885	4.679	0.000	19.030
i	0	170	36	6.041	6.816	5.402	4.649	0.000	25.000
i	1	81	13	6.334	7.276	5.468	4.283	0.000	17.810
we	0	170	36	0.181	0.251	0.120	0.451	0.000	4.320
we	1	81	13	0.249	0.366	0.142	0.502	0.000	3.110
you	0	170	36	0.377	0.470	0.293	0.598	0.000	3.300
you	1	81	13	0.491	0.692	0.302	0.963	0.000	6.520
shehe	0	170	36	0.443	0.585	0.314	0.987	0.000	7.490
shehe	1	81	13	0.329	0.478	0.208	0.612	0.000	2.610
they	0	170	36	0.399	0.494	0.320	0.572	0.000	3.230
they	1	81	13	0.433	0.561	0.326	0.589	0.000	3.020
ipron	0	170	36	5.262	5.706	4.839	3.172	0.000	16.220
ipron	1	81	13	5.623	6.343	4.949	3.170	0.000	13.850
article	0	170	36	10.755	11.532	10.000	5.214	0.000	50.000
article	1	81	13	10.217	11.328	9.129	5.309	0.000	32.310
prep	0	170	36	13.244	13.895	12.547	4.448	0.000	26.030
prep	1	81	13	13.113	14.066	12.035	4.708	0.000	23.640
auxverb	0	170	36	7.539	8.100	6.866	4.130	0.000	15.530
auxverb	1	81	13	8.183	8.955	7.355	3.833	0.000	15.290
adverb	0	170	36	5.801	6.298	5.265	3.507	0.000	16.670
adverb	1	81	13	6.143	6.904	5.534	3.171	0.000	15.620
conj	0	170	36	6.011	6.564	5.458	3.395	0.000	14.810
conj	1	81	13	6.490	7.165	5.863	2.918	0.000	13.020
negate	0	170	36	1.205	1.451	0.997	1.535	0.000	14.290
negate	1	81	13	1.288	1.510	1.060	1.032	0.000	4.810
verb	0	170	36	16.097	16.948	15.224	5.755	0.000	34.620
verb	1	81	13	16.864	18.107	15.475	5.801	0.000	28.120
adj	0	170	36	4.555	5.004	4.150	3.058	0.000	29.410
adj	1	81	13	4.364	4.859	3.891	2.222	0.000	13.330
compare	0	170	36	4.007	4.472	3.527	3.270	0.000	18.550
compare	1	81	13	4.280	5.025	3.541	3.576	0.000	14.550
interrog	0	170	36	0.686	0.795	0.577	0.727	0.000	3.880
interrog	1	81	13	0.846	1.014	0.680	0.778	0.000	3.120
number	0	170	36	1.290	2.566	0.570	7.719	0.000	100.000
number	1	81	13	0.707	1.037	0.476	1.334	0.000	11.320
quant	0	170	36	1.807	2.029	1.603	1.377	0.000	8.330
quant	1	81	13	1.847	2.144	1.575	1.314	0.000	7.260
affect	0	170	36	2.412	2.755	2.104	2.209	0.000	17.650
affect	1	81	13	2.472	2.827	2.158	1.523	0.000	6.330
posemo	0	170	36	1.612	1.866	1.377	1.591	0.000	11.760
posemo	1	81	13	1.654	1.953	1.374	1.265	0.000	6.110

					95% CI Mean				
	Hit	**Valid**	**Missing**	***M***	***UL***	***LL***	***SD***	***Min***	***Max***
negemo	0	170	36	0.790	0.970	0.629	1.092	0.000	6.580
negemo	1	81	13	0.805	0.977	0.625	0.826	0.000	4.260
anx	0	170	36	0.193	0.265	0.133	0.425	0.000	2.780
anx	1	81	13	0.247	0.360	0.148	0.480	0.000	2.440
anger	0	170	36	0.156	0.207	0.110	0.327	0.000	2.630
anger	1	81	13	0.131	0.204	0.072	0.308	0.000	2.130
sad	0	170	36	0.214	0.317	0.123	0.658	0.000	5.880
sad	1	81	13	0.190	0.255	0.129	0.299	0.000	1.540
social	0	170	36	4.031	4.450	3.623	2.761	0.000	14.290
social	1	81	13	4.236	4.884	3.668	2.743	0.000	15.540
family	0	170	36	0.162	0.218	0.115	0.331	0.000	1.950
family	1	81	13	0.252	0.372	0.150	0.521	0.000	2.610
friend	0	170	36	0.157	0.221	0.103	0.397	0.000	3.450
friend	1	81	13	0.157	0.246	0.087	0.349	0.000	2.430
female	0	170	36	0.402	0.579	0.268	1.051	0.000	10.000
female	1	81	13	0.383	0.526	0.256	0.607	0.000	2.550
male	0	170	36	0.513	0.677	0.383	0.994	0.000	8.440
male	1	81	13	0.403	0.548	0.273	0.604	0.000	3.580
cogproc	0	170	36	9.928	10.685	9.047	5.537	0.000	25.000
cogproc	1	81	13	11.281	12.331	10.210	4.993	0.000	22.340
insight	0	170	36	2.549	2.861	2.242	2.095	0.000	11.110
insight	1	81	13	2.733	3.153	2.344	1.816	0.000	6.830
cause	0	170	36	0.651	0.763	0.543	0.708	0.000	3.860
cause	1	81	13	0.810	0.978	0.649	0.769	0.000	4.740
discrep	0	170	36	0.554	0.688	0.441	0.878	0.000	6.250
discrep	1	81	13	0.674	0.826	0.521	0.705	0.000	3.180
tentat	0	170	36	3.857	4.240	3.458	2.688	0.000	14.190
tentat	1	81	13	4.194	4.788	3.542	2.847	0.000	16.020
certain	0	170	36	0.655	0.775	0.543	0.795	0.000	6.250
certain	1	81	13	0.757	0.908	0.618	0.663	0.000	2.420
differ	0	170	36	2.712	3.029	2.409	2.056	0.000	8.680
differ	1	81	13	3.378	3.782	2.993	1.988	0.000	9.380
percept	0	170	36	8.851	9.746	8.045	5.526	0.000	41.180
percept	1	81	13	8.416	9.153	7.602	3.639	0.000	20.000
see	0	170	36	5.456	6.147	4.859	4.413	0.000	34.550
see	1	81	13	5.327	6.152	4.563	3.573	0.000	20.000
hear	0	170	36	1.029	1.366	0.765	2.014	0.000	22.220
hear	1	81	13	0.695	0.861	0.525	0.739	0.000	2.850
feel	0	170	36	2.114	2.593	1.738	2.712	0.000	23.530
feel	1	81	13	2.104	2.591	1.647	2.213	0.000	14.290
bio	0	170	36	3.349	3.864	2.879	3.351	0.000	27.590

					95% CI Mean				
	Hit	**Valid**	**Missing**	***M***	***UL***	***LL***	***SD***	***Min***	***Max***
bio	1	81	13	3.179	3.944	2.503	3.602	0.000	28.570
body	0	170	36	1.801	2.128	1.473	2.434	0.000	17.240
body	1	81	13	1.579	1.951	1.260	1.582	0.000	6.920
health	0	170	36	0.395	0.566	0.260	1.050	0.000	10.340
health	1	81	13	0.452	0.871	0.209	1.604	0.000	14.290
sexual	0	170	36	0.019	0.034	0.008	0.084	0.000	0.730
sexual	1	81	13	0.009	0.019	0.001	0.042	0.000	0.260
ingest	0	170	36	1.196	1.472	0.946	1.768	0.000	11.760
ingest	1	81	13	1.115	1.526	0.731	1.979	0.000	14.290
drives	0	170	36	3.149	3.456	2.846	2.047	0.000	10.110
drives	1	81	13	3.410	3.809	3.025	1.876	0.000	8.240
affiliation	0	170	36	0.606	0.750	0.479	0.886	0.000	5.760
affiliation	1	81	13	0.799	1.024	0.585	0.986	0.000	4.020
achieve	0	170	36	0.418	0.504	0.337	0.554	0.000	2.610
achieve	1	81	13	0.439	0.543	0.344	0.467	0.000	2.130
power	0	170	36	1.596	1.788	1.397	1.374	0.000	6.820
power	1	81	13	1.599	1.857	1.356	1.151	0.000	5.210
reward	0	170	36	0.539	0.659	0.431	0.731	0.000	4.100
reward	1	81	13	0.589	0.723	0.463	0.568	0.000	2.530
risk	0	170	36	0.163	0.223	0.109	0.388	0.000	2.780
risk	1	81	13	0.148	0.214	0.094	0.270	0.000	1.380
focuspast	0	170	36	1.556	1.773	1.339	1.439	0.000	8.330
focuspast	1	81	13	1.975	2.367	1.621	1.687	0.000	6.900
focuspresent	0	170	36	11.996	12.825	11.160	5.663	0.000	25.000
focuspresent	1	81	13	12.317	13.502	11.069	5.620	0.000	22.290
focusfuture	0	170	36	0.850	0.994	0.721	0.892	0.000	6.150
focusfuture	1	81	13	1.045	1.239	0.848	0.867	0.000	3.650
relativ	0	170	36	14.900	15.647	14.097	5.164	0.000	29.790
relativ	1	81	13	15.322	16.291	14.421	4.040	7.550	26.470
motion	0	170	36	2.804	3.103	2.506	2.023	0.000	15.170
motion	1	81	13	2.933	3.274	2.550	1.715	0.000	8.450
space	0	170	36	9.108	9.760	8.544	4.002	0.000	23.530
space	1	81	13	8.948	9.676	8.279	3.236	0.000	18.380
time	0	170	36	3.332	3.619	3.055	1.895	0.000	9.710
time	1	81	13	3.801	4.487	3.301	2.625	0.000	19.050
work	0	170	36	0.525	0.637	0.412	0.798	0.000	4.790
work	1	81	13	0.586	0.796	0.411	0.888	0.000	5.660
leisure	0	170	36	2.058	2.399	1.745	2.236	0.000	14.290
leisure	1	81	13	1.639	2.083	1.295	1.823	0.000	12.680
home	0	170	36	0.764	0.926	0.613	1.097	0.000	7.580
home	1	81	13	1.010	1.506	0.646	1.937	0.000	14.290

	Hit	Valid	Missing	*M*	95% CI Mean *UL*	95% CI Mean *LL*	*SD*	*Min*	*Max*
money	0	170	36	0.117	0.161	0.075	0.314	0.000	2.200
money	1	81	13	0.099	0.155	0.057	0.219	0.000	1.330
relig	0	170	36	0.165	0.236	0.109	0.422	0.000	3.450
relig	1	81	13	0.189	0.286	0.114	0.397	0.000	2.520
death	0	170	36	0.099	0.140	0.062	0.269	0.000	1.930
death	1	81	13	0.132	0.217	0.072	0.342	0.000	2.520
informal	0	170	36	0.719	0.938	0.529	1.459	0.000	11.540
informal	1	81	13	0.634	0.935	0.397	1.294	0.000	9.520
swear	0	170	36	0.038	0.077	0.009	0.242	0.000	2.630
swear	1	81	13	0.032	0.060	0.009	0.122	0.000	0.840
netspeak	0	170	36	0.050	0.091	0.020	0.225	0.000	1.750
netspeak	1	81	13	0.052	0.082	0.028	0.127	0.000	0.660
assent	0	170	36	0.209	0.285	0.142	0.491	0.000	3.340
assent	1	81	13	0.156	0.259	0.078	0.414	0.000	2.840
nonflu	0	170	36	0.422	0.634	0.251	1.363	0.000	11.540
nonflu	1	81	13	0.387	0.650	0.195	1.132	0.000	9.520
filler	0	170	36	0.011	0.022	0.003	0.066	0.000	0.790
filler	1	81	13	0.008	0.017	0.002	0.034	0.000	0.250

Note. The Hit column denotes session outcome where 0 = miss, 1 = hit. 95% *CIs* for mean estimated with 1,000 bootstraps.

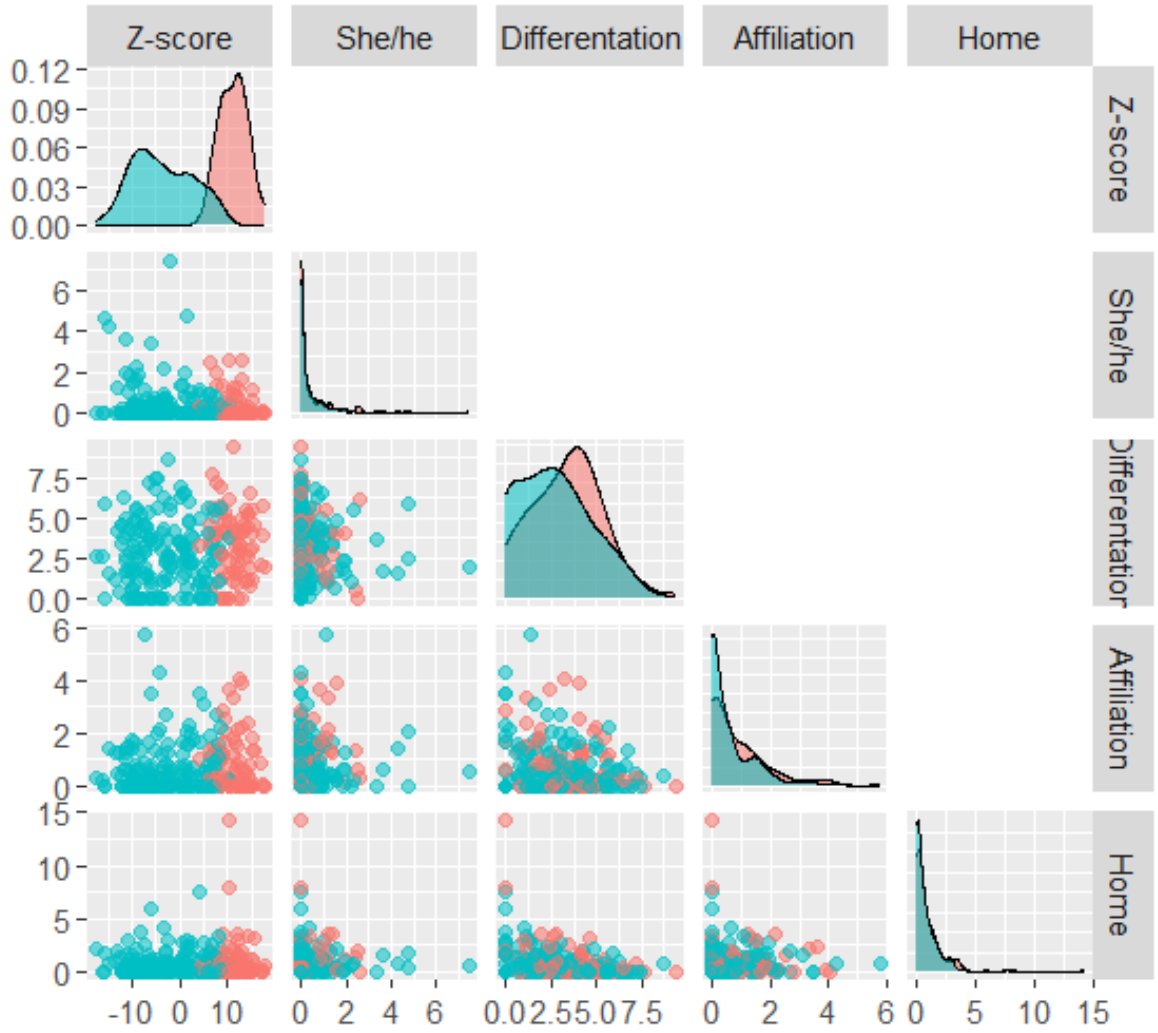
Supplemental Table 3

Model Summary statistics for Session z-score as outcome

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	*p*
1	0.000	0.000	0.000	8.854	0.000		0	247	
2	0.129	0.017	0.013	8.798	0.017	4.169	1	246	0.042
3	0.190	0.036	0.028	8.728	0.019	4.934	1	245	0.027
4	0.230	0.053	0.041	8.669	0.017	4.364	1	244	0.038
5	0.267	0.071	0.056	8.603	0.018	4.729	1	243	0.031

Supplemental Figure 1

Pairs Plot of Variables from Stepwise Model Selection with z-score as outcome



Note. Pink represents Hit and blue represents Miss.

Supplemental Table 4

Model Summary statistics for Binary Hit Rate as outcome

Model	Deviance	AIC	BIC	df	ΔX²	*p*	McFadden R²	Nagelkerke R²	Tjur R²	Cox & Snell R²
1	315.705	317.705	321.231	250			0.000		0.000	
2	309.929	313.929	320.980	249	5.776	0.016	0.018	0.032	0.023	0.023
3	305.572	311.572	322.149	248	4.357	0.037	0.032	0.055	0.039	0.040
4	301.733	309.733	323.835	247	3.839	0.050	0.044	0.076	0.054	0.054
5	299.401	309.401	327.028	246	2.333	0.127	0.052	0.088	0.064	0.063
6	297.145	309.145	330.298	245	2.255	0.133	0.059	0.100	0.072	0.071

Supplemental Figure 2

Pairs Plot of Variables from Stepwise Model Selection with Binary Hit Rate as outcome

Note. Pink represents Hit and blue represents Miss.